

82-1
6-8-70





Bureau Educational Research
DAVID H. ... COLLEGE
Dated 6-8-70
Ass. No. 927



A NOTE ON SIR CYRIL BURT'S 'FACTORIAL ANALYSIS OF QUALITATIVE DATA'

By LOUIS GUTTMAN

Scientific Director, Israel Institute of Applied Social Research, Jerusalem

I. *Procedures for Dealing with Qualitative Data.* II. *Analysis into Principal Components.* III. *Applications to Scalable and Non-Scalable Data.* IV. *The Study of Deviations : Image Analysis and Nodal Analysis.*

I. PROCEDURES FOR DEALING WITH QUALITATIVE DATA

World War II interrupted communication between scientists in different countries, and only gradually is the exchange of information being made up. A case in point is the monograph *The Prediction of Personal Adjustment* by Paul Horst and others, which appeared just before the United States actively entered the holocaust. This was published by the Social Science Research Council in New York as *Bulletin 48* in 1941. During the war, only a handful of copies was able to reach Europe. In a recent article in this *Journal* (1), Sir Cyril Burt has called attention to the importance of developing a theory for qualitative data, as distinct from quantitative data. He develops a particular algebraic formulation which leads to the resolution of the data into principal components (latent vectors). This happens to be a topic treated also in the above-mentioned monograph, by the present writer. It is gratifying to see how Professor Burt has independently arrived at much the same formulation. This convergence of thinking lends credence to the suitability of the approach. The purpose of the present note is to call attention to the points of similarity and to describe further developments which have occurred in the United States and in Israel in the theory of qualitative data.

In a chapter of the above-mentioned Social Science Research Council monograph entitled 'The Quantification of a Class of Attributes' (2), it was proposed that qualitative data could be recorded in a manner amenable to treatment by matrix algebra. In form, the matrix M on page 326 of the monograph is identical with Table I of Professor Burt's article ((1), p. 171). My own paper proposes three different kinds of problems of quantification :

(a) Deriving quantitative *scores* for the respondents, to maximize a certain correlation ratio (for the *rows* of the matrix).

(b) Deriving quantitative *weights* for the categories of the qualitative items, to maximize another correlation ratio (for the *columns* of the matrix).

(c) Deriving scores and weights simultaneously, to maximize a certain correlation coefficient (for rows and columns of the matrix simultaneously).

It was proved that all three problems are essentially equivalent. The maximizing scores and weights for the third problem are also the maximizing ones for the

first and second problems respectively. Furthermore, the maximum correlation ratios and coefficient are always mutually equal for the same data.

II. ANALYSIS INTO PRINCIPAL COMPONENTS

As usual in analysing internal consistency, the equations sought lead not only to a *maximum*, but also to a series of other stationary values down to a minimum. These various solutions provide a principal component system. There is a set of principal component scores for the respondents, and a set of principal component weights for the categories of the items. Professor Burt has arrived at the latter solution, namely, that for principal component weights.

My own article goes on to point out that, while the principal components here are formally similar to those for quantitative variables, nevertheless their interpretation may be quite different. The interrelations among qualitative items are not linear, nor even algebraic, in general. Similarly, the relation of a qualitative item to a quantitative variable is in general non-algebraic. Since the purpose of principal components—or any other method of factor analysis—is to help reproduce the original data, one must take into account this peculiar feature.

The first principal component can possibly fully reproduce all the qualitative items entirely by itself: the items may be perfect, albeit non-algebraic, functions of this component. *Linear* prediction will not be perfect in this case, but this is not the best prediction technique possible for such data. Therefore, if the first principal component only accounts for a small proportion of the total variance of the data in the ordinary sense, it must be remembered that this ordinary sense implies *linear* prediction. If the correct, but *non-linear*, prediction technique is used, the whole variation can sometimes be accounted for by but the single component. In such a case, the existence of more than one principal component arises merely from the fact that a linear system is being used to approximate a non-linear one. (Each item is always a perfect linear function of *all* the principal components taken simultaneously.)

III. APPLICATIONS TO SCALABLE AND NON-SCALABLE DATA

This thinking was carried further with the definition of a perfect scale (3). In a perfect scale, each item is a perfect function of a single rank order of respondents. The functional relation is not algebraic, although of an extremely simple nature. Applying the general equations of principal components developed in (2) to the case of the perfect scale defined in (3) leads to some most extraordinary results. There is a definite *law of formation* for the entire set of principal components for scores, and a corresponding law of formation for principal component weights. Indeed, the matrix equation for the principal components in each case transforms into a *difference* equation. For dichotomous items, this is a second order linear difference equation which, in the limit, for an infinite number of items, yields a second order *differential* equation that is classical in mathematics and physics. The principal components turned out to be sets of orthogonal functions, like the Legendre polynomials, Hermite polynomials, trigonometric functions, etc.

The second principal component in all cases is a perfect U- or J-shaped function of the first. The third principal component is also a perfect function of the first, but its graph shows a curve with two bends (for example, a cubic parabola). The

linear correlation between any two principal components is always zero, because they are orthogonal to each other (and have zero means); but the *curvilinear* correlation of each with the first component is perfect. In this sense, the further principal components beyond the first add no new information.

In another sense, it has been found that the further principal components provide a *psychological* frame of reference for scalable data. The second principal component for scores has been identified for many attitudinal data as the *intensity* of attitude (5). More recently, the third and the fourth component for scores have been found to portray somewhat new psychological concepts which have been named *closure* and *involution*, respectively (9).

For non-scalable data, the principal components will not, of course, have a law of formation such as that just described, and indeed may have no useful law at all. Progress has also been made in developing meaningful structural theories for non-scalable qualitative data, and here various new kinds of systems of principal components appear possible. For the beginning of these developments, which go under the general name of the 'theory of facets,' see (10).

An unexpected consequence of these discoveries for qualitative data has been the rethinking it has stimulated about factor analysis in general. A new approach has been developed for qualitative data with linear regressions, which also leads to laws of formation of principal components. The general theory is called that of the *radex* and it contains the *simplex* and *circumplex* as subtheories. Empirical data tend to confirm this new approach for the case of tests of mental abilities (11).

IV. THE STUDY OF DEVIATIONS: IMAGE ANALYSIS AND NODAL ANALYSIS

The above-mentioned type of difference equation, and its lawful principal components, holds for the case of a perfect scale. An important problem that must be considered arises from the fact that perfect scales are rarely, if ever, to be found in practice. More typical may be the case of quasi-scales, or scales 'plus' error. Here, then, it becomes important to distinguish between the structural and the deviant aspects of the observed data: only the former will have a law of formation.

Two new general approaches for dealing with the problem of deviations from perfect structures are called *image analysis* and *nodal analysis* respectively. Both of these are useful for many kinds of non-scale structures as well as for quasi-scales.

One image technique, Israel^{*} Alpha, for studying quasi-scales has been described in (14). The Israel Beta technique has also been used successfully, and will be described in future publications.

Nodal analysis can be regarded as a generalization and simplification of Lazarsfeld's *latent structure* theory. It leads to easy computing routines, whether for quasi-scale or non-scale structures (15).

Here we cannot do more than to call attention to these newer developments, most of which have not yet been published in much detail. The work done thus far in probing the possibilities of structural theories for qualitative data seems to indicate that we are merely scratching at the beginnings of a rich and intricate subject, which fully deserves to be studied in its own right.

REFERENCES

1. Burt, Cyril (1950). 'The factorial analysis of qualitative data.' *Brit. J. Psych., Stat. Sect.*, III, 166-85.
2. Guttman, Louis (1941). 'The quantification of a class of attributes,' in Horst, P., *et al.* *The Prediction of Personal Adjustment*. New York : Social Science Research Council, 319-48.
3. Guttman, Louis (1944). 'A basis for scaling qualitative data.' *American Sociological Review*, IX, 139-50.
4. Guttman, Louis (1947). 'The Cornell technique for scale and intensity analysis.' *Educ. and Psychol. Measurement*, VII, 247-80.
5. Guttman, Louis, and Suchman, E. A. (1947). 'Intensity and a zero point for attitude analysis.' *Amer. Sociol. Rev.*, XII, 57-67.
6. Guttman, Louis (1947). 'Scale and intensity analysis for attitude, opinion, and achievement,' in Kelly, George A. (editor). *New Methods in Applied Psychology*. Maryland : University of Maryland Press, 173-80.
7. Guttman, Louis (1950). 'The principal components of scale analysis,' ch. 9 in Stouffer, S. A., *et al.* *Measurement and Prediction*. Princeton : University Press.
8. Guttman, Louis (1951). 'Scale analysis, factor analysis, and Dr. Eysenck.' *Internat. J. Opinion and Attitude Research*, 5, 103-20.
9. Guttman, Louis. 'The principal components of scalable attitudes,' in Lazarsfeld, Paul F. (editor). *Mathematical Thinking in the Social Sciences*. New York : Columbia University Press (in the press).
10. Guttman, Louis. 'The theory of facets.' *Ibid.*
11. Guttman, Louis. 'A new approach to factor analysis : the radex.' *Ibid.*
12. Guttman, Louis, and Suchman, E. A. (1947). 'A solution to the problem of question "bias."' *Public Opinion Quarterly*, XI, 445-55.
13. Guttman, Louis, and Foa, Uriel G. (1951). 'Social contact and an intergroup attitude.' *Public Opinion Quarterly*, XV, 43-53.
14. Guttman, Louis (1951). *The Israel Alpha technique for scale analysis* (stencilled). Available on request from the Israel Institute of Applied Social Research.
15. Guttman, Louis. 'The theory of nodal analysis' (in preparation).

SCALE ANALYSIS AND FACTOR ANALYSIS

*Comments on Dr. Guttman's Paper*¹

By CYRIL BURT

University College, London

I. *The Analysis of Answer Patterns.* II. *Criticisms of the Factorial Approach.* III. *Factor Analysis of a Typical Scale Pattern.* IV. *Factor Analysis of the Corresponding Answer Pattern.* V. *Summary.*

I. THE ANALYSIS OF ANSWER PATTERNS

Quantitative and Qualitative Data. It is encouraging to find that Dr. Guttman has discerned points of resemblance between the methods we have independently reached for analysing qualitative data; and I willingly agree that this convergence in our lines of approach lends additional plausibility to the results. If, as I gather, he cannot wholly accept my own interpretations, that perhaps is attributable to the fact that our starting-points were rather different. My aim was to factorize such data; his to construct a scale. In the paper² to which he has referred, I sought to develop a practicable procedure for discovering the factors underlying qualitative characteristics, and then to apply that procedure to certain recurrent problems of my own. Dr. Guttman's purpose, as the title of his previous contribution indicates, was to present 'a theory and method of scale construction' by means of 'quantifying a class of attributes.' I have myself commented³ on the similarity between my equations for factor-measurements and the equations for component scores given by Dr. Guttman, and ascribed it to the fact that his mode of deriving components seemed identical with the method proposed by Pearson for calculating 'index characters'—a method which had formed the original basis for my own factorial procedures.

Dr. Guttman's technique of scale analysis has been developed primarily in connexion with "the study of the internal structure of a universe of attitude and opinion items"; and the data he has used to illustrate his procedure consist, as a rule, of responses to questionnaires. I myself, however, first encountered the type of problem which he has discussed in a rather different field, namely, in the attempt to refine the scaling of the original Binet-Simon intelligence-tests. Before the publication of Binet's 1911 version, British psychologists relied chiefly on tests of an 'internally graded' type—opposites, dotting, card sorting, and the like. Binet introduced a composite *échelle* comprising a number of 'externally graded items,' i.e., questions or tasks for which the responses differ not in quantity but only in quality, and which are simply marked 'right' or 'wrong.'⁴ Thus,

¹ I am much indebted to Dr. Guttman for his kindness in reading my manuscript and to Dr. M. L. Fraser for his help in checking my calculations.

² This *Journal*, III, pp. 166–85. As there explained, the theoretical part of the paper was based on a technical memorandum prepared during the early stages of the war for the Department of Personnel Selection of the War Office. The factorial procedure had already been applied to qualitative physical characteristics, and to data obtained from the Binet-Simon Scale, for which it was originally developed.

³ This *Journal*, V, p. 200. Dr. Guttman observes (p. 2 above) that the method I have followed yields the same solution as his for the 'principal component weights.' But, as I shall show in a moment, there is (with a suitable standardization of the data) a still more striking resemblance between the figures reached by the two different methods for the 'principal component scores.'

⁴ For a discussion of the principles underlying the two types of test, see Burt, C. (1914), 'The Measurement of Intelligence by the Binet Tests,' *Eugen. Rev.*, VI, pp. 36–50, 140–52.

with an internally graded test the numerical data consist of measurements; with the externally graded, of frequencies. Hence for the former the appropriate statistical methods appeared to be those developed for 'quantitative variables' (e.g., the ordinary product-moment correlation); for the latter, those described by Yule for what he called 'attributes' (e.g., methods of contingency and association).

In the chapters contributed to *Measurement and Prediction* (18) both Dr. Guttman and Dr. Lazarsfeld draw a sharp distinction between the principles involved in these two cases. Factor analysis, they maintain, has been elaborated solely with reference to data which is quantitative *ab initio*; hence, they suppose, it cannot be suitably applied to qualitative data. On this side of the Atlantic, however, there has always been a tendency to treat the two cases together, and, with this double application in view, to define the relevant functions in such a way that they will (so far as possible) cover both simultaneously.¹ British factorists, without specifying very precisely the assumptions involved, have used much the same procedures for either type of material. Nevertheless, there must of necessity be certain minor differences in the detailed treatment. These were briefly indicated in the paper Dr. Guttman has cited; but they evidently call for a closer examination. I think in the end it will be found that they are much slighter than might be supposed.

Example. With qualitative material, the special problems that arise and the general mode of approach can be best explained if we start with a concrete illustration. Suppose we want to classify the children in some particular age-group—all those aged six, let us say—into 5 grades of ability by means of a series of dichotomous test-items like the questions in the Binet Scale. If each item constituted a pure and reliable test for the ability to be assessed, then, arguing along the lines of 'information theory,'² and proceeding by a progressive subclassification, we should evidently need six items altogether: namely, (i) a test like 5 Numbers, which would separate those below average from those above, i.e., those with a mental age of less than 6 from those with a mental age of 6 or more; (ii) a test like 4 Numbers, which would divide those with a mental age below average into those with a mental age of less than 5, and those with a mental age of 5 to 6; and (iii) a test like 3 Numbers, which would divide those with a mental age of less than 5 into those with a mental age of less than 4, and those with a mental age of 4 to 5. Similarly, we should require three more tests to subclassify those who were above the average for age 6.

In an experiment planned to study the working of the tests, the data obtained would be entered on a mark sheet similar to that shown in Table I. This is a special case of the more general table³ printed in my previous article ((17), Table I). Now, however, each 'determinable' presents two 'determinates' only, i.e., each item yields two 'categories', namely, 'pass' (marked 1) and 'fail' (marked 0). British investigators⁴ describe such tables as 'answer patterns'; and, from the early days of mental testing, they have devoted a good deal of attention to the simpler types of pattern which the answers from these non-graded tests appear to display.

Table I records the results obtained with six tests from the London version of the Binet Scale, when applied to a representative sample of 6-year-old school children, from whom, for simplicity, I have chosen 10 individuals only. It may be compared with the typical table printed by Dr. Guttman in his discussion of the 'components of scale analysis' ((18), p. 317, Table I).

The method of recording differs from his in two minor respects. First of all, in accordance with the principle followed by most practical test-constructors, the battery is designed to contain at least one item so easy that all the children can answer it and at least one so difficult that all are

¹ See, for example, (14), I, pp. 15f., where the Stieltjes integral is used to facilitate this combined treatment. The work of the French school (Borel and his colleagues), by deriving the theory of measurement from the theory of sets, makes the relations still simpler: (see (15), opening chapters). A yet more recent and relevant example is Kendall's 'general coefficient of correlation' which applies equally well to measurements, ranks, and dichotomous classifications (cf. Kendall, M. G., *Rank Correlation Methods*, 1948, pp. 17f.).

² Cf. Burt, C., this *Journal*, IV, p. 196. The principles involved are analogous to, but of course not identical with, the conversion of an integral measurement from the 'common scale of notation' to the 'binary' where only two digits, 0 and 1, are used.

³ Dr. Guttman says in his paper (p. 1 above) that this table is identical in form with "the matrix *M* on p. 326 of the monograph." I should have thought it was rather to be compared with the very first table in his monograph (p. 321), since that (like mine) presents the initial record sheet which we require to analyse, and shows the various items explicitly classified into categories and subcategories.

⁴ The same mode of tabulation is adopted by Ferguson and other British writers: cf. (6) and (11).

TABLE I. ANSWER PATTERN FOR DICHOTOMOUS TESTS

Tests	Persons										Total	Differences
	A.A.	B.A.	B.B.	C.A.	C.B.	C.C.	C.D.	D.A.	D.B.	E.A.		
Repeating—												
3 Numbers..	1	1	1	1	1	1	1	1	1	1	10	1
4 Numbers..	0	1	1	1	1	1	1	1	1	1	9	2
5 Numbers..	0	0	0	1	1	1	1	1	1	1	7	4
6 Numbers..	0	0	0	0	0	0	0	1	1	1	3	2
7 Numbers..	0	0	0	0	0	0	0	0	0	1	1	1
8 Numbers..	0	0	0	0	0	0	0	0	0	0	0	
Total ..	1	2	2	3	3	3	3	4	4	5	30	
Deviates ..	-2	-1	-1	0	0	0	0	+1	+1	+2		
Frequencies ..	1	2			4			2		1	10	

likely to fail. Hence $N + 1$ tests are deemed necessary to discriminate N types of persons. For theoretical work, however, where we are more interested in factorizing items than in grading persons, it is better to include $n + 1$ types of persons in order to analyse the answers from n tests. Secondly, the ordinary answer pattern gives only the positive replies, whereas Dr. Guttman enters negative as well as positive, and thus has $2n$ categories for n tests. It will be convenient to keep his phrase—'scale pattern'—to designate his type of table, and to reserve the phrase 'answer pattern' to designate a mark sheet of the more familiar type.

Problem. In the earlier investigations of the Binet and similar tests, one of my chief objects was to determine the factors in terms of which such answer patterns could best be described. The factors appeared to be of two kinds—'content factors' and 'difficulty factors.'¹ Could we so purify our scale that it was virtually a perfect instrument for measuring 'general intelligence,' all factors resulting from special abilities would be eliminated. In such a case there would still remain two obvious factors for which we should wish to secure specifications—the general factor of 'difficulty,' common to all the tests, and the general factor of 'ability,' common to all the persons.

In practice the former is usually obtained by correlating persons, the latter by correlating tests. If we insist on calculating correlations, we must begin by reducing the crude scores to deviations. But if this preliminary reduction is omitted, both factors can be obtained from the same matrix by factorizing rows for the one and columns for the other. With dichotomous items, however, like those illustrated in Table I, provided we are content to give each person and each item an equal weight, we can avoid the more elaborate calculations required for a full and formal analysis, and assess the ability of the children by simply adding the number of correct replies obtained for each item, and the difficulty of the items by simply adding the number of correct replies given by each child: (cf. the marginal totals in Table I). As the answer pattern above plainly indicates, there is a certain reciprocity between the two factors; and, indeed, either concept may be defined in terms of the other.²

The Unique Answer Pattern. When the sample of items and the sample of persons have been picked more or less at random, then the frequency-distributions for both will be approximately normal or at least bell-shaped. In practice, however, an attempt is generally made to choose items that will yield a rectilinear scale. Binet's method of standardizing his problems by grouping children according to chronological age was intended to provide a rectilinear distribution for persons as well (or rather for categories of persons). Evidently, if we had one perfectly reliable item for each level of difficulty and one perfectly representative child for each level of ability, the resulting answer pattern would then be uniquely determined.

Let us consider the effect of such a double selection of items and persons on an answer pattern like that set out above. When the items have been efficiently selected, but the children are unselected (or just randomly selected from each age-group), then the line dividing successes from failures will approximate to a normal ogive. In such a case, if the number of children is large, the table showing the answer pattern is usually condensed so as to allot a column to each age-group

¹ The 'difficulty' factors found in such analyses have been criticized by Mr. Gourlay as artefacts. For his criticisms and a reply, see this *Journal*, IV, pp. 65–76.

² Cf. Burt, C., 'Test Construction and the Scaling of Items,' this *Journal*, IV, 1951, p. 106, Sect. III (a).

instead of to each child, and percentages are usually entered in place of separate units; but these percentages still exhibit an approximation to the ogival form (cf. (4), pp. 132-3, Table III, and (19), pp. 113f.). When both the tests and the persons have been appropriately selected, the dividing line should be perfectly straight, and the area containing the successful answers a right-angled triangle.

From this it will be clear that the peculiar type of answer pattern which would be given by a test comprised of a perfectly consistent set of items would possess several distinctive peculiarities.¹ (i) If with such a pattern of answers we compute an order of difficulty for the test-items from any sub-group of persons within the sample, that order should always remain the same, no matter how the sub-group is selected. (ii) The N_i persons passing any given item will always be the N_i most able persons in the group, no matter what N_i may be. (iii) A total of x marks will always be obtained by passing the x least difficult items in the series, no matter what x may be. It follows (iv) that the frequencies in the distribution of persons (shown near the bottom of the table) will be identical with the first differences between the successive numbers of persons passing each test (shown at the side of the table). Thus, as Walker points out (6), except so far as there may be ties for individuals having the same level of ability, or for items belonging to the same level of difficulty, the answer pattern will be completely fixed by the order of the items and the order of the persons. If, therefore, the frequency distribution of the persons is stated, we can deduce from the total mark gained by any one person, not only how many items, but also which items, he has in fact passed.

The Effects of Inconsistency. With the majority of the pass-or-fail tests in common use, the order of difficulty is far from stable; and the resulting pattern never displays the perfect regularity just described. The Binet tests, for example, are intended to measure a hypothetical factor called general intelligence: how far is it safe to assume that all the items consistently measure this factor and nothing else? With the six memory tests enumerated above, any child who answered a harder test would almost certainly answer all the easier tests, and any child who failed in an easier test would almost certainly fail in all the harder. But now suppose we add Binet's question requiring the child to give the date. A bright child of 6 who scores a mental age of 8 will probably fail in the new test but succeed in repeating 6 numbers; a dull child of 10 with the same mental age will probably fail with 6 numbers but succeed in the new test. Omit '6 Numbers,' and the second will seem the more intelligent; omit 'Giving the Date,' and the first will score most. Hence the verdicts of the two tests are no longer consistent; and, largely for this reason, Binet dropped both from the final version of his scale.²

For such irregularities there are two main causes. Both persons and items suffer from what might be called an internal unreliability. To some extent this can be obliterated by grouping. We may, for example, begin by grouping persons according to age, ability, sex, social status, and the like. Even then we usually find that the order varies for different groups: (cf. the tables of rank-differences in (4), pp. 143, 194). When they are large enough to be statistically significant, such variations imply attainments of the persons and the special content factors of the tests. For certain statistical purposes, the differences can be still further smoothed out by grouping the items. But for practical purposes it is better to recognize their existence, and regard them as a cue for improving the way the items have been selected. Binet, indeed, in his successive revisions attempted to purify his scale by eliminating most of those that appeared to be affected by scholastic attainments or by special aptitudes. However, he neither used nor possessed any objective method for determining which items were relatively pure and which were not.

A theoretical study of the structures exhibited by perfect and imperfect scales shows that irregularities of the type described must inevitably lower the efficiency of the test-series as a whole, and explains just how such irregularities operate. In particular it is found that they tend (a) to diminish the validity of the separate items, as assessed by the correlation of each with the total marks obtained from all, and (b) to render the composite test less reliable, less valid, and (since they inevitably reduce the variance of the marks) less discriminative (cf. (11), pp. 55-8). It therefore becomes desirable to find some method for measuring what I have called the 'homogeneity' and 'heterogeneity' of any test-battery proposed for practical use.

¹ Ferguson's account of the properties of the 'unique answer pattern matrix' ((11), p. 53) seems incomplete. As an earlier critic has pointed out, "he begins with 'firstly,' but then either he (or the printer) appears to have omitted certain characteristic properties, specified by other writers." Nor can I agree with Ferguson's suggestion that the "unique configuration" furnished by a perfect scale "seems to indicate that the mind has a certain structure" (p. 59): it results from the way the investigator has constructed his composite test.

² A number of similar inconsistencies between the responses obtained with the original version of the scale were pointed out in an early report. The most conspicuous occurred with semi-scholastic items like Transcription and Reading—tasks which nearly every normal child learns to perform at a fairly definite age in his school life, whether he is slightly above or slightly below the average level of intelligence. The irregularities in the actual answer patterns obtained from the earlier surveys furnished the first and most obvious indications of the presence of such unsatisfactory items: (cf. (4), pp. 132-4, 146f.).

The Measurement of Heterogeneity. The general characteristics of consistent and inconsistent answer patterns will now be clear. As we have seen, when the tests are arranged in order of difficulty and the persons in order of ability, then the answer pattern for an ideal test should display a solid block of 1's above the diagonal line of division (i.e., in what I shall call the 'success triangle') and a solid block of 0's below it (in what I shall call the 'failure triangle'). Inconsistencies will be indicated by erratic 0's appearing in the former area and erratic 1's in the latter; and the pattern will in consequence no longer show a sharp or cleanly defined line separating all the individual passes from all the individual failures. The effect of these irregularities Thomson designated by the term 'higgledypiggledyness.' Any coefficient obtained by counting up their number (or some equivalent procedure) he proposed to call a 'coefficient of hig' and its complement a 'coefficient of unig' (uniqueness). The words he has coined avoid the ambiguities of vaguer terms, like 'heterogeneity' and 'inconsistency,' which are used in so many contexts with so many different meanings.

With a perfectly haphazard distribution of answers we should expect half the total number of successes to be scattered over one triangle, and half over the other. An obvious index of homogeneity or 'unig' (U) is therefore given by the formula

$$U = \frac{\text{No. of Actual Successes in Success Triangle} - \frac{1}{2} \text{ Total No. of Successes}}{\frac{1}{2} \text{ Total No. of Successes}}$$

Writing this in the form

$$U = 1 - \frac{4 \times \text{No. of Successes in Failure Triangle}}{\text{Total No. of Responses}} = 1 - \frac{4 \sum e_f}{n \times N},$$

we see that it is virtually the same as Guttman's 'coefficient of reproducibility' ((18), p. 77).

In practice, however, with the ordinary type of composite test, such a coefficient does not prove altogether satisfactory. For one thing, with the same set of items, it varies according to the distribution of the sample of persons. In such a case an alternative method is to compare the frequencies for the individuals with the first differences between the totals for the tests; and accordingly, after considering various possibilities, Walker proposed a 'coefficient of hig' based on the sum of the squares of the discrepancies between the two sets of figures. But even this, as he himself points out in a more recent paper, is not entirely free from objection (6).

There are several other ways in which a coefficient might be conceivably developed. We might, for instance, use any of the coefficients of intrinsic reliability, e.g., the split half method, or the variance method which I have described as giving the mean value for all possible split halves.¹ Or again we might construct from the data the ideal answer pattern which gives the closest fit, and test the discrepancies by the square contingency ('chi-squared'); the agreement could then be measured by the root mean square contingency ϕ . When the scales or questionnaires include a large number of items, such procedures become decidedly laborious; and probably most British writers would concur with Ferguson's conclusion: "at present no convenient measure is available for estimating the divergence of an obtained answer-pattern matrix from a theoretically unique matrix" ((11) p. 54).

The Advantages of a Factorial Analysis. For general purposes, the best course, in my own view, is to undertake a factorial analysis of the data obtained. Since we are interested primarily in difficulty factors, the natural approach would be to correlate persons. If the persons are grouped in categories, the matrix of correlations or product-sums need not be large. Indeed, it will often be sufficient to compare the rank-orders of the several items for the different groups ((4), pp. 132f., 146f.). For more intensive studies, a complete analysis is desirable,² particularly as that may incidentally shed light on the presence of

¹ Burt, C., 'The Reliability of Teachers' Assessments,' *Brit. J. Educ. Psychol.*, XV, 1945, pp. 91-2. The formula giving the mean for all possible split-halves appears to be essentially the same as that given by Kendall for the 'coefficient of concordance' ((14), p. 411).

² This is especially true where the data relate to non-cognitive characteristics, e.g., social attitudes, Rorschach answers, and the like. Many of the items in a case-history are as a rule recorded only in terms of presence or absence—e.g., the occurrence of some accident, disease, or hypothetical gene, and especially certain recurrent features in the social environment—parents or relatives of this or that type, membership of this or that group or circle, contact with this or that member of the group. In all such cases the resulting 'nought-and-one' record-table can be treated like a 'measurement matrix' and factorized in the manner described; and the results obtained often yield suggestive ways of describing the intricate pattern of a life history or a social group (cf. (17), p. 166 and refs.).

irrelevant content factors. In such a case it will be better to start by correlating or comultiplying items rather than persons. The method of weighted summation, with unreduced item-variances, should be employed by preference. But simple summation will furnish a good approximation. The working procedures, and the results obtained in the case of empirical test-items, have been sufficiently discussed and illustrated in earlier papers (cf. (13), (15), (17), and refs.). In view of recent criticisms, however, it seems desirable to consider more closely how far it may be legitimate to treat qualitative data in this fashion, and particularly whether such methods can yield satisfactory results in the case of a perfect scale.

II. CRITICISMS OF THE FACTORIAL APPROACH

The Relations between Scale Analysis and Factor Analysis. In his very instructive paper (pp. 1-4) Dr. Guttman has indicated what he takes to be the differences, as well as the similarities, between his approach and my own. He observes that, while the 'principal components' employed in his method of scale analysis may be 'formally similar' to those employed in factor analysis, "nevertheless their interpretation may be quite different." The differences are examined more fully in his earlier contribution on 'The Relation of Scalogram Analysis to Other Techniques.'¹

If I understand Dr. Guttman rightly, his objections turn on five main points.

1. First, he holds that factor analysis is "designed only for quantitative variables," and is consequently unsuited for qualitative data ((18), p. 191). The method, he says, "originated as a single factor theory by Spearman, and was developed into a multifactor theory by Thurstone and others." Were his own procedure to be described in factorial terms, then, he adds, we should have to treat it as "a single factor theory for qualitative data."

However, as other writers have pointed out (cf. this *Journal*, V, p. 206), such statements limit the term 'factor analysis' to very specific forms. In point of fact, the technique now generally known as factor analysis is much older than Spearman's procedures. It originated with Pearson's proposal to reduce a given multivariate distribution to terms of the 'principal axes of the frequency ellipsoid' and take these axes as representing 'index characters.' It was Pearson's investigation of the general problem that really supplied the earliest "algebraic formulation, leading" (if I may borrow Dr. Guttman's phrase) "to the resolution of the data into principal components (latent vectors)" ((2), pp. 559f.; cf. (15), p. 309). Factor analysis was thus a multifactor method from the start. Spearman's single factor method was developed several years later as a substitute, because he held that Pearson's approach was unsuited to the data obtained in psychology. Nevertheless, in spite of his criticisms, numerous researches were carried out in which Pearson's method was applied, not only to quantitative measurements, such as those furnished by graded tests, but also to qualitative data, such as were supplied by dichotomous test-problems and by questionnaires. In a footnote, Dr. Guttman ((18), p. 193) refers to the treatment worked out by Yule (Karl Pearson's assistant) for dealing with frequency-tables for qualitative variables (3), and considers it "strange that statistical text-books in the social sciences have not followed suit, but fail to discuss material of this kind at all." But as a matter of fact in this country psychologists have made free use of Yule's procedures, especially in relation to social data²; and numerous theses could be cited where factorial methods have been applied to tables of frequencies, contingencies, or Yule's coefficients of colligation or point-correlation. In particular, 'answer patterns' have regularly been subjected to a factorial analysis by various devices (cf. (6), p. 326, (11), p. 52, and refs.).

¹ ((18), pp. 191f.; in the same volume Dr. Lazarsfeld also brings forward somewhat similar criticisms; cf. *ibid.*, pp. 469f.).

² In dealing with the item-responses obtained with the Binet Scale (where the classification was in terms of pass or fail) and with many social conditions affecting delinquency (where the classification was in terms of presence or absence) Yule's formulæ were used at a very early stage (cf. *Mental and Scholastic Tests*, 1921, pp. 217f., and *The Young Delinquent*, 1925, pp. 54f.). The tetrachoric coefficient advocated by Pearson and criticized by so many recent writers came into vogue somewhat later (cf. (13)). Yule's method of treating frequency-tables in cases of manifold classification involves fitting the matrix of observed frequencies with a matrix of expected frequencies derived by a procedure identical with that of centroid analysis. And when, instead of the more laborious method of 'weighted summation' entailed by Pearson's equations, I put forward the formula for 'simple summation'—the formula afterwards termed the 'centroid formula' by Thurstone—I pointed out that it was, in effect, an extension to quantitative variables of the usual formula for investigating association in the case of qualitative attributes (cf. (3), p. 64, eqns. (1) and (2)).

2. Secondly, Dr. Guttman argues that, in order to apply factor analysis, we must begin by calculating correlation coefficients, and that in the case of qualitative data such coefficients are bound to be misleading. With his criticisms of the uses made of the tetrachoric coefficient and the point correlation I very largely agree. Yet his arguments seem only to prove that these coefficients are not suitable for *all* occasions or for *every* purpose. There is one coefficient which he does not explicitly discuss—the ordinary product-moment coefficient applied to the data after they have been transformed to standard measure; and this, which has been frequently used for such problems, yields (as I shall show in a moment) results that are virtually the same as his. Nevertheless, nothing in the theory of factor analysis confines its application solely to coefficients of correlation. Indeed, for the factorist there is often a special advantage in working with frequencies, since (as I pointed out in the paper to which Dr. Guttman refers) the higher order frequencies may prove especially serviceable in the calculation of group factors.¹

3. His third argument runs as follows. The principal criterion for scalability is reproducibility. But factor analysis does not allow us to reproduce the original data from the so-called factor-measurements. Hence factor analysis can never show whether a scale is perfect or not. This objection, however, applies only to the special procedures advocated by Spearman and Thurstone, which seek to analyse, not the total variance, but merely the common factor-variance. The method of principal axes, on the other hand, requires the full test-variances to be retained in the covariance matrix; and with Pearson's procedure the factor-measurements are obtained by pre-multiplying the initial measurements by the matrix of direction cosines (the latent vectors). Now such a matrix is necessarily orthogonal. Hence its transpose can be used as a second pre-multiplier to reproduce the initial measurements from the factor-measurements (cf. (10), Appendix II). An exact reproduction is therefore possible. If, then, we can also show that, with a perfect scale, one of the factors so obtained is in perfect correlation with the rank of the persons, it would seem that the method can after all provide an entirely satisfactory criterion.

4. "The Spearman-Thurstone approach to factor analysis," as Dr. Guttman says, "is completely linear, and is therefore not adequate for analysing the curvilinearities inherent in the scale pattern." Certainly in that mode of approach the factor-measurements are always estimated by the method of linear regression, as developed in Pearson's earlier papers. But Pearson himself also elaborated a method for dealing with curvilinear regressions.² His treatment was intended primarily for problems involving an external criterion; but it is equally applicable to the case of an internal criterion or factor. Elsewhere I have argued that it is quite unnecessary to restrict the theory of factor analysis to linear relations only ((10), p. 258); and, with the aid of the orthogonal polynomials used in the theory of curvilinear regression,³ it is simple to estimate factor-measurements from test data or test data from factor-measurements on the assumption of non-linear relations.

5. Finally, Dr. Guttman concludes that "from a scale analysis it can be known what a factor analysis will show; from a factor analysis it will usually be difficult, if not impossible, to know what a scale analysis will show." To determine this point I propose to apply a factorial procedure to his own table, and see how far the results achieved are similar to those reached by his own scale analysis. A concrete numerical example will probably help best to explain how far our methods are similar and in what ways they seem to differ.

III. A FACTOR ANALYSIS OF A TYPICAL SCALE PATTERN

An Answer Pattern for Five Items. In Table II I have reproduced the 'scale pattern' which Dr. Guttman uses in his chapter on 'The Principal Components of Scale Analysis' ((18), pp. 317, 333) to illustrate his own procedure. It represents a 'perfect scale' formed by the responses of 6 'types of person' to five test-items classified under 10 'types of category': (for brevity I shall refer to them simply as 'persons' and 'categories'). In the last line I have appended the value to which the entries should be changed in order to secure the same square-sum (1.000) for each category.

¹ (17), p. 181. The memorandum already quoted gave a worked example from the Binet tests, and an algebraic proof (summarized in this *Journal*, V, p. 202).

² The earliest systematic study of curvilinear regression is to be found in Pearson, K., 'On the General Theory of Skew Correlation and Non-Linear Regression,' *Biometric Series*, II, 1905; cf. also *id.*, 'On a General Method of Determining the Successive Terms in a Skew Regression Line,' *Biometrika*, XIII, 1921, pp. 296f.

³ See (14), II (esp. pp. 158f., 'Curvilinear Regression: Case when the Independent Variate proceeds by Equal Steps,' and refs.), (5), and (8). Greenleaf (7) also includes useful refs.

Scale Analysis and Factor Analysis

This table corresponds to Table I in my previous article ((17), p. 171). There, however, the possibility was envisaged of more than two categories for each item, and those referring to the same item were kept together. Here there are two categories only—positive (e.g., correct answers) marked *a*, and negative (e.g., incorrect answers) marked *b*. Adopting Dr. Guttman's mode of tabulation, I have now grouped all categories of the same type together, regardless of item, and have numbered the items from 1 to 5 as their difficulty decreases, and the persons from 5 to 0 as their ability decreases. The reader may perhaps find it easier to interpret the factorial procedure if he thinks of the items as five test-problems from an ideal version of the Binet Scale, applied to typical children of six successive mental ages.

TABLE II. SCALE PATTERN FOR FIVE DICHOTOMOUS ITEMS (GUTTMAN)

Persons	Categories										Frequency of Response
	1a	2a	3a	4a	5a	1b	2b	3b	4b	5b	
5	1	1	1	1	1	0	0	0	0	0	5
4	0	1	1	1	1	1	0	0	0	0	5
3	0	0	1	1	1	1	1	0	0	0	5
2	0	0	0	1	1	1	1	1	0	0	5
1	0	0	0	0	1	1	1	1	1	0	5
0	0	0	0	0	0	1	1	1	1	1	5
Frequency of Response	1	2	3	4	5	5	4	3	2	1	30
Standardized Value ($1/\sqrt{N_j}$)	1	$\frac{1}{\sqrt{2}}$	$\frac{1}{\sqrt{3}}$	$\frac{1}{\sqrt{4}}$	$\frac{1}{\sqrt{5}}$	$\frac{1}{\sqrt{5}}$	$\frac{1}{\sqrt{4}}$	$\frac{1}{\sqrt{3}}$	$\frac{1}{\sqrt{2}}$	1	

Factor Analysis. The method of weighted summation involves analysing the symmetric matrix of standardized product-sums into its latent roots and latent vectors (cf. Burt, *Factors of the Mind*, Appendix II). Accordingly we begin by calculating a symmetric contingency table, i.e., the matrix of product-sums for items, $C_i = M_i M_i'$ ((17), p. 172, eqn. (i)); and then standardize¹ it in the usual way by pre- and post-multiplying it by a diagonal matrix, $D^{-\frac{1}{2}}$, where the elements of $D^{\frac{1}{2}}$ are $\sqrt{N_j}$ (where N_j denotes the number of persons in each category). The figures thus obtained are shown in Table III: (to save space I omit the last 4 rows, since they are obvious from the fact that the complete table must be axisymmetric).

TABLE III. STANDARDIZED MATRIX OF PRODUCT-SUMS (*M*)

Category	1a	2a	3a	4a	5a	1b	2b	3b	4b	5b
1a	1.000	.707	.577	.500	.447	.000	.000	.000	.000	.000
2a	.707	1.000	.816	.707	.632	.316	.000	.000	.000	.000
3a	.577	.816	1.000	.866	.775	.516	.289	.000	.000	.000
4a	.500	.707	.866	1.000	.894	.671	.500	.289	.000	.000
5a	.447	.632	.775	.894	1.000	.800	.671	.516	.316	.000
1b	.000	.316	.516	.671	.800	1.000	.894	.775	.632	.477
..	..	..	..	..	..	..	..	..	..	..

(a) *Factor-Saturations.* The 'saturations' for the *i*th factor have been computed in the usual way by multiplying the normalized elements specifying the *i*th latent vector by the square root of the *i*th latent root. Six factors are obtained; and the resulting matrix is orthogonal by columns (Table IV).

¹As explained in my previous article, the main object of standardizing the data is to weight the frequencies so as to compensate for the differences in numbers in the various sub-categories. Here an incidental effect is to weight each item according to its difficulty.

TABLE IV. FACTOR-SATURATIONS FOR CATEGORIES (F)

Types of Category	Factor	G	BI	BII	BIII	BIV	BV	Square-Sum
1a		+·408	+·598	+·545	+·373	+·189	+·063	1·000
2a		+·577	+·676	+·309	—·105	—·267	—·178	1·000
3a		+·707	+·621	·000	—·258	·000	+·218	1·000
4a		+·816	+·478	—·218	—·074	+·189	—·126	1·000
5a		+·913	+·267	—·244	+·167	—·085	+·028	1·000
1b		+·913	—·267	—·244	—·167	—·085	—·028	1·000
..		..	..	..	..	..	..	..
Square-Sum (v)		5·000	3·000	1·000	·500	·300	·200	10·000

(b) *Factor-Measurements.* To obtain the partial regressions or 'weights' for the several categories ($W' = V^{-1}F'$) it is only necessary to divide each column of factor-saturations by the square-sum shown at the foot. We then apply these weights to the standardized form of the initial matrix of marks. The weighted sums yield the values for the several factor-measurements ($P = W'M$, where $M = D^{-\frac{1}{2}}M_1$). These are given in Table V.

As I have endeavoured to show in the Appendix, where the scale pattern is perfect, the factor-measurements can be obtained more directly by the formulæ used in calculating the constants for curvilinear regressions. These figures have already been tabulated. Thus if the reader will turn to the tables of orthogonal polynomials given by Fisher and Yates,¹ and take the values of ξ_i for $i = 1, 2, \dots, 5$, normalizing each column by the aid of the square-sums printed at the foot, he will find that the figures so obtained are identical with the factor-measurements given in Table V (except for the first factor which always has the value $+1/\sqrt{n}$). This set of tables may be used for a battery containing any number of items from 2 to 75.

Taking the factor-saturations as weights, it will be found that we can reconstruct the original standardized measurements *exactly* by the usual equation $M = FP$. With this method of factor analysis, therefore, it can no longer be objected that the initial data are not reproducible from the factorial results.

TABLE V. FACTOR-MEASUREMENTS FOR PERSONS (P)

Type of Person	5	4	3	2	1	0	Square-Sum
Factor G	+·408	+·408	+·408	+·408	+·408	+·408	1·000
" BI	+·598	+·359	+·120	—·120	—·359	—·598	1·000
" BII	+·545	—·109	—·436	—·436	—·109	+·545	1·000
" BIII	+·373	—·522	—·298	+·298	+·522	—·373	1·000
" BIV	+·189	—·567	+·378	+·378	—·567	+·189	1·000
" BV	+·063	—·315	+·630	—·630	+·315	—·063	1·000
Square-Sum	1·000	1·000	1·000	1·000	1·000	1·000	6·000

Scale Analysis. Dr. Guttman's 'component weights and scores' are derived directly from the initial table of data. He computes no coefficients of correlation or contingency, nor does he print any matrix of product-sums. The 'component weights' (if required) he calculates only *after* the 'component scores' have been obtained. His threefold theoretical approach has been succinctly summarized in the preceding paper (p. 1 above); and the algebraic equations thus obtained (given in his previous publications) bring to light a number of new and interesting features. Yet I am doubtful whether, for the ordinary student,

¹ (16), Table XXIII, p. 70. The analogy between this mode of approach and that adopted by Fisher in fitting curved regression lines suggests that an adaptation of his treatment might be used to test significance in the case of empirical data: but that is a question I hope to take up in a later paper.

the working procedure which the equations suggest would be easier to follow or quicker to apply than the ordinary factorial procedure, especially as that can be appreciably simplified in the case of an artificially perfect scale.

(a) *Component Scores.* The table given by Dr. Guttman as obtained by his own procedure is reproduced in Table VI. Like the factor-measurements given in Table V, the scores given by Dr. Guttman are orthogonal by factors (i.e., the rows are uncorrelated); but, unlike the former, they are not orthogonal by persons. His figures in fact are identical with those to be found in the table printed by Fisher and Yates¹ ((16), p. 70).

TABLE VI. SCORES FOR PERSONS (GUTTMAN)

Component	Persons						Square-Sum
	5	4	3	2	1	0	
Constant	1	1	1	1	1	1	6
I	5	3	1	-1	-3	-5	70
II	5	-1	-4	-4	-1	5	84
III	5	-7	-4	4	7	-5	180
IV	1	-3	2	2	-3	1	28
V	1	-5	10	-10	5	-1	252

(b) *Component Weights.* Having dealt with the problem of "quantifying the rank order of people," Dr. Guttman turns to that of "quantifying the category values of items." These values he terms the 'component weights.' The figures that he gives ((18), Table 5, p. 329) are not proportional to my figures either for the 'factor-saturations' or for the 'regression weights.' They are, however, indirectly related. We can in fact derive the proportionate values for Dr. Guttman's 'weights' by dividing the saturations for each category (given in Table IV) by the square root of the total number of positive entries for that category, that is, multiplying them by $1/\sqrt{N}$. We then arrive at the figures shown in Table VII. The effect of this procedure is to allot the same weight for the general factor to every category in the list, while, with the remaining factors, the figures obtained are now simple multiples of the smallest figure in each column. In Dr. Guttman's table of 'principal component weights' the figures are either identical with these multiples or with simple fractions of them.

TABLE VII. UNSTANDARDIZED FACTOR-SATURATIONS (U)

Category	Multiplier	G	BI	BII	BIII	BIV	BV
1a	$1/\sqrt{1} = 1.000$	.408	.598	.545	.373	.189	.063
2a	$1/\sqrt{2} = .707$	.408	.478	.218	-.075	-.189	-.126
3a	$1/\sqrt{3} = .577$	.408	.359	.000	.149	.000	.126
4a	$1/\sqrt{4} = .500$	.408	.239	-.109	-.037	.094	-.063
5a	$1/\sqrt{5} = .447$	.408	.119	-.109	.075	.038	.012
1b	$1/\sqrt{5} = .447$	.408	-.119	-.109	-.075	-.038	-.012
..	..	..	..	..	..	..	..

He insists that his use of the conventional term 'weights' should not be taken to imply that such figures are needed to secure summational equations for computing the persons' scores. Nevertheless, as he remarks, "it so happens that the equations of internal consistency prove that there is a *reciprocal relation* between principal components for people and principal components for categories that does involve an additive (more precisely, an *averaging*) process, so we are justified in part in retaining the traditional nomenclature of 'scores' and 'weights.'" This is in effect a recognition of what I have elsewhere called the 'reciprocity principle.'²

¹ As Fisher and Yates point out, the figures are proportionate values only, and for convenience the figures given by the algebraic formulæ have been converted to whole numbers by the fractions shown at the foot of each table. As it happens, a table identical with Dr. Guttman's is also printed in the analysis of an example used by Mather to illustrate curvilinear regression (*Statistical Analysis in Biology*, 1942, p. 139). The student will find the detailed discussion in Mather's chapter exceedingly illuminating, particularly in regard to the analysis of variance.

² Cf. (10), pp. 487-94. In the notation there used, writing $F = LV\frac{1}{2}$, we have $M = LV\frac{1}{2}P$, where $LL' = I$ and $PP' = I$. Hence $P = V^{-\frac{1}{2}}L'M$ and $L = MP'V^{-\frac{1}{2}}$. Thus L can be regarded either as normalized factor-saturations (weights) for tests obtained by comultiplying tests or as normalized factor-measurements for tests obtained by comultiplying persons (in that case using $P'V^{-\frac{1}{2}}$ as weights). Similarly there will be two alternative interpretations for P .

could, if we preferred, substitute Yule's triangular mode of factorization : this would show still more explicitly that each of the five bipolar factors dealt with five progressive levels of difficulty. I need not exemplify this point in detail here, since I have already illustrated it in a previous contribution.¹

IV. FACTOR ANALYSIS OF THE CORRESPONDING ANSWER PATTERN

Tabulation of Positive Categories Only. I now propose to consider the results that would have been obtained had we taken the experimental data embodied in Dr. Guttman's table and treated them according to the more usual routine methods. There will be two differences at the very outset. First, with two types of answer only, the actual performances would ordinarily be entered in a table specifying only 5 items, not 10 categories. Hence the rank of the product-sum matrix will be only 5, and we shall now obtain only 5 factors. Secondly, for purposes of correlation, the entries would be converted into deviation form. Since we have assumed $n = N - 1$, this reduction will make the number of degrees of freedom for persons the same as that for tests, and the effect of the change will be to weight both successful and unsuccessful answers according to the difficulty of the items. The resulting mark sheet is shown in Table VIII.

TABLE VIII. MARK SHEET FOR DR. GUTTMAN'S DATA IN UNITARY STANDARD MEASURE

Item	Persons						Square-Sum
	5	4	3	2	1	0	
1	5	-1	-1	-1	-1	-1	6 × 5
2	4	4	-2	-2	-2	-2	6 × 8
3	3	3	3	-3	-3	-3	6 × 9
4	2	2	2	2	-4	-4	6 × 8
5	1	1	1	1	1	-5	6 × 5
Total	15	9	3	-3	-9	-15	6 × 105

The 5×5 correlation table can be factorized by weighted summation in the usual way. The saturations and regressions thus obtained are set out in Tables IXA and IXB.

Factor-Saturations. The process of converting the observed test-marks to deviations about the mean has eliminated the former general factor (G), which was really a factor of difficulty for the several test-items ; and the new first factor now represents the general ability used in solving all such items. The factor-measurements for this factor therefore will measure the general ability of the persons tested.

The relation between the two sets of saturations will be clear if we compare the figures in Tables VII and IXA. Let us drop the first column in Table VII, and then restandardize the remaining figures so that the square-sum is still 1.000. This is equivalent to multiplying them by the values shown in the second column of Table IX (headed 'multiplier') ; and yields the results shown in the table. The sign pattern remains the same ; and the number of oscillations in the five successive rows is now 0, 1, 2, 3, and 4 respectively.

Factor-Measurements. To find the ordinary linear regression coefficients, each column of saturations is, as usual, divided by its square-sum. Then, on applying these coefficients (Table IXB) to the standardized test-measurements (obtained by dividing the marks in Table VIII by the square root of the square-sum), we obtain the factor-measurements for the five factors. It will be discovered that the figures thus computed are precisely the same as those obtained by the previous procedure (Table V above). They are, therefore, exactly proportional to Dr. Guttman's 'scores' ; and differ from his only by being expressed in unitary standard measure. Hence it may be fairly claimed that, so far as the essential results are concerned, the figures reached by an ordinary factorial analysis are identical with those reached by scale analysis.

¹ This *Journal*, IV, p. 76, Table IIB ; cf. also *Factors of the Mind*, p. 306, Table IIIB.

TABLE IX. FACTOR-SATURATIONS AND REGRESSIONS

Item	Multi- plier	A. Saturations					B. Regressions				
		Factor					Factor				
		I	II	III	IV	V	I	II	III	IV	V
1	$\sqrt{6/5}$	·6547	+ ·5976	+ ·4083	+ ·2073	+ ·0690	·2182	+ ·5976	+ ·8166	+ ·6908	+ ·3451
2	$\sqrt{6/4}$	·8281	+ ·3780	- ·1291	- ·3272	- ·2182	·2760	+ ·3780	- ·2582	- 1·0906	- 1·0911
3	$\sqrt{6/3}$	·8783	+ ·0000	- ·3651	+ ·0000	+ ·3086	·2928	+ ·0000	- ·7302	+ ·0000	+ 1·5430
4	$\sqrt{6/2}$	·8281	- ·3780	- ·1291	+ ·3273	- ·2182	·2760	- ·3780	- ·2582	+ 1·0906	- 1·0911
5	$\sqrt{6/1}$	·6547	- ·5976	+ ·4083	- ·2073	+ ·0690	·2182	- ·5976	+ ·8166	- ·6908	+ ·3451
Square-Sum		3·0000	1·0000	·5000	·3000	·2000	·3333	1·0000	2·0000	3·3333	5·0000

Factor Analysis by Simple Averaging. In virtue of the symmetry that now characterizes each column of Table IX, the manner in which both saturations and regressions effect a hierarchical classification is here a little more obvious. Nevertheless, a set of 5 items and a sample of 6 persons do not lend themselves very neatly to classification by means of successive dichotomies. In order to exhibit the general principle in its most typical form we require an example with $n = 2^k$ items. Let us therefore drop the third item in Dr. Guttman's set of items. This will leave two persons in the same category, but may be all the more instructive on that account: (see top section of Table X).

The oldest device for securing what are essentially factor-measurements for the 'general factor' common to all the tests in a battery consists in just averaging the marks gained by each person. This is equivalent to multiplying the matrix of marks by a vector of weights which has every element positive and equal to $1/n$. When the number of items is an integral power of 2, the rest of the factor-measurements may similarly be obtained by multiplying the same matrix by vectors, each of which has every element numerically equal to $1/n$, but differs in sign pattern so that all the vectors have zero correlations with one another. If in every vector except the first the number of positive signs is equal to the number of negative, it will be possible to construct n such vectors for a battery of n items. On arranging them in order of increasing complexity, we shall find the demand for non-successive factor. The whole procedure is tantamount to taking as the first set of factor-measurements the simple averages of the marks; then averaging the residuals; then averaging the residuals from this second set of averages; and so on. It is in fact the method most commonly adopted for partitioning a given set of data in the analysis of variance.¹

To keep the square-sum for factors equal to that for items, it will be necessary to normalize the vectors by substituting $1/\sqrt{n}$ for $1/n$: see left hand panel of the middle section of Table X. On pre-multiplying the marks for the four remaining tests by this orthogonal matrix of weights we obtain the factor-measurements shown in the central panel. To recover the initial matrix of marks from the factor-measurements we have merely to take the transpose of the 'item weights,' and use it to pre-multiply the matrix of factor-measurements, as shown in the last panel of all.

Now let us compare the results thus obtained with the values in Tables IV, V, and IX (ignoring the item and the factor which have been suppressed). It will be seen that the 'item weights' and 'factor weights' in Table X have the same sign pattern as the regressions and saturations in the preceding tables, and that (except for the indeterminate sign of the zeros) the 'factor-measurements' also have the same sign pattern as the 'factor-measurements' in Table V, while their numerical values furnish a plausible approximation to the 'best fitting values' given by the more elaborate procedure.

¹ See Mather, *loc. cit. sup.*, pp. 61-4 (esp. sect. 23, 'The Principles of Partition'). Cf. also Burt, C., 'Factor Analysis and Analysis of Variance,' this *Journal*, I, 1947, pp. 5f. The factor-measurements thus computed are linearly independent, but the first is not statistically independent of all the others: nor do the figures give the closest possible fit. In order that the first condition may be fulfilled, we should weight each set of factor-measurements appropriately before subtracting them to obtain successive residuals. For this modified method of obtaining factor-measurements direct from test-measurements, see Burt, C., 'The Influence of Differential Weighting,' this *Journal*, III, 1950, esp. pp. 115f.: it yields the same measurements as are obtained by simple summation (centroid formula with unreduced communalities). If the second condition is essential, the 'latent vectors' must be computed: these give the closest fit as judged by least squares.

TABLE X. FACTOR ANALYSIS BY SIMPLE AVERAGING

Items	Persons						Square-Sum				
	5	4	3	2	1	0					
1	5	-1	-1	-1	-1	-1	30				
2	4	4	-2	-2	-2	-2	48				
4	2	2	2	2	-4	-4	48				
5	1	1	1	1	1	-5	30				
	Item Weights				Factor-Measurements						
I	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	6	3	0	0	-3	-6	90
II	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	3	0	-3	-3	0	3	36
III	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	0	-3	0	0	3	0	18
IV	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	1	-2	1	1	-2	1	12
	Factor Weights				Initial Scores (Reconstructed)						
	I	II	III	IV							
1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	5	-1	-1	-1	-1	-1	30
2	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	$-\frac{1}{2}$	4	4	-2	-2	-2	-2	48
4	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	2	2	2	2	-4	-4	48
5	$\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	1	1	1	1	1	-5	30

The mode of classifying both items and persons implicitly adopted by summational methods of factorization will now be clearer. The first of the four factors—the general factor—has positive values for each of the ‘item weights’: when the values are equal, the effect is virtually to sum the correct responses; it thus specifies, in Dr. Guttman’s phrase, the ‘metric’ component. The second factor contrasts performance in the two harder items with performance in the two easier: in a certain sense, therefore, it is, as he says, an ‘intensity component.’ The third factor contrasts performance in the extreme items with performance in those of intermediate difficulty: it thus corresponds with Dr. Guttman’s component of ‘closure.’ If a name were needed for the fourth factor, I would call it an ‘alternating’ component: Dr. Guttman believes that it ‘portrays a new psychological concept’ which he terms ‘involution.’

Since the number of oscillations that characterizes each successive factor is primarily an outcome of the mathematical technique employed, it would seem better to envisage two ways of labelling the factors obtained. (1) First, it would be useful to have purely abstract designations merely indicating the *formal* properties of the several factors. And these designations would presumably differ according to our formal interpretation of the procedure. We may, for example, conceive the factors either as specifying the means (weighted or unweighted) of the observed measurements and of their successive residuals or as specifying the several types of independent comparison. (2) Secondly, we need concrete designations giving the *material* interpretation of the factors in this or that particular application; and the interpretation of course will vary according to the subject matter to which the analysis is applied.¹ From the formal standpoint (which alone concerns us here) the whole procedure, as I have suggested elsewhere, is rather like analysing the sounds of a musical instrument of complex timbre into a fundamental note with its successive overtones. The ideal test might be compared to a relatively pure note with little or no distorting noise; an analysis by subdivided factors is comparable to a Fourier analysis, where each additional harmonic is fainter than its predecessor and has twice the number of oscillations per second; and the recombination of factor-measurements by weighted summation is analogous to the compounding of simple harmonic motions at right angles to each other, as demonstrated (in the two-dimensional case) by Lissajous figures.²

¹ Spearman and Thurstone appear to use the word ‘factor’ exclusively in the latter sense. Elsewhere I have suggested that the term ‘component’ might be used for the purely mathematical ‘factor,’ and the name ‘factor’ reserved for the concrete interpretation.

² A still closer analogy is presented by the methods adopted for the ‘spectral analysis of aggregates’ in quantum physics; and, in view of the use of complex numbers in quantum theory, it is instructive to note that some of the problems arising in the factor analysis of frequency-data may also be solved by the use of components involving complex numbers (cf. Burt, C., ‘The Unit Hierarchy and its Properties,’ *Psychometrika*, III, 1938, pp. 151f.; also this *Journal*, IV, 1951, p. 20). In observing how the attempt to factorize a pattern of all-or-none responses gives rise to a set of oscillating components, it is impossible not to be reminded of the way de Broglie and Schrödinger have attempted to describe electrons in terms of oscillations in multidimensional space.

Conclusions. The foregoing examples will, I hope, sufficiently indicate that, with ideal tables of the type we have just been considering, a factor analysis can after all "show what a scale analysis will show," and that the use of the ordinary product-moment coefficient may lead to results which are by no means "misleading," even when the scale is perfect: a more general proof is outlined in the Appendix. But further it may, I think, also be claimed that a factorial approach such as I have described is particularly suitable with empirical data where the test-items do not form a perfect scale. This conclusion, it must be confessed, is based chiefly on experience in factorizing dichotomous test-items (like those composing such scales as the Binet-Simon or group-tests of the ordinary type) and qualitative assessments obtained for physical characteristics and for social conditions.

Now, as I emphasized at the outset, the problems discussed by Dr. Guttman originated in a rather different field. "The need for scale analysis," he tells us, "arises out of the fundamental problem of attitude scale and opinion polling of how to determine the dimensions of meaning which the questions asked have for the respondents: scale analysis affords a rigorous test for the existence of single-meaning for an area, and provides a rank order of individuals for such areas as are found scalable" ((18), p. 88). My own impression is that each of the various methods we have discussed has its special advantages and its special limitations, and the chief question awaiting solution is to determine still more precisely the particular conditions which give the preference to this method or to that. With this aim in view, it would, for instance, be instructive to study the results reached by applying the factorial methods described above to data obtained from questionnaires on attitudes or opinions, and by using Dr. Guttman's methods of scalogram analysis for data obtained with ordinary pass-or-fail test-items. Meanwhile, there can be no doubt that his intensive examination of the problems encountered in attempts to "quantify attributes" has brought to light many novel and instructive points, and should, in his phrase, go far to "stimulate rethinking about factor analysis in general."

V. SUMMARY

1. In analysing qualitative data, like those obtained from questionnaires or from tests where the answers are simply marked right or wrong, the numerical processes that arise are primarily those associated with counting rather than with measurement. In such cases, therefore, the appropriate statistical methods will be based on the principles worked out by Pearson and Yule for 'characters not quantitatively measurable' (to use Pearson's description) or 'attributes' (as Yule more succinctly termed them). Nevertheless, whether we are dealing with qualitative data or with quantitative, the problems that present themselves, and the procedures that are available, turn out in the main to be closely analogous. The study of scaling techniques, of total, partial, and multiple dependence, and of hypothetical factors or components, follows lines that are very much the same in either case.

2. As a preliminary to scaling the items that compose a questionnaire or test, the investigation of the unique type of 'answer pattern' that would be furnished by an ideal or perfect scale has proved highly illuminating, and has furnished various plausible indices for measuring the degree to which an empirical scale approximates to the pure or perfect type, e.g., the coefficients of 'hig' and 'unig' discussed by Walker, Ferguson, and others. The simplest of these has the same purpose and the same general character as the 'coefficient of reducibility' proposed by Dr. Guttman. Nevertheless, Dr. Guttman's more recent and intensive examination of the issues raised has elucidated much that was previously obscure, and revealed many new and fruitful problems.

3. Attempts to improve the construction and scaling of tests like the Binet-Simon series appear to indicate that for general purposes the most satisfactory method consists in a factorial analysis of the constituent items. It is applicable to empirical, as well as to ideal or perfect, scales. The results indicate a distinction between 'content factors' and 'difficulty factors,' and demonstrate a reciprocal relation between the difficulty of the test-items and the abilities of the persons tested.

4. For the study of qualitative data the most effective factorial technique would appear to be that originally proposed by Pearson for determining the orthogonal mean square regression lines (i.e., weighted summation with unreduced test-variances) or, as Pearson

himself preferred to call it, the method of 'principal axes.' Pearson's extension of regression to the determination of curvilinear regression lines suggests a generalization of factorial methods to deal with cases in which the relations are not linear but curvilinear. These modified methods are applicable both to measurement-data and to frequency-data. The use of regression equations of a degree higher than the first is analogous to the use of higher moments, and, with qualitative data, leads to the consideration of frequencies of higher order than the second. With ideal answer patterns the resulting factor-measurements (the latent vectors for persons) are proportional to the values tabulated by Fisher and Yates for orthogonal polynomials. It would seem therefore that Dr. Guttman's objections to the use of factor analysis with qualitative data may thus be fully met.

5. A factor analysis of Dr. Guttman's ideal 'scale pattern' (which includes negative as well as positive categories) yields figures for the factor-measurements which are identical with Dr. Guttman's 'component scores' when the latter are reduced to normalized form. The factor-saturations are related to his 'component weights,' but here the relation is less direct.

6. If an answer pattern of the more usual type (i.e., one which includes positive categories only) is substituted for the ideal 'scale pattern,' a factor analysis of the product-moment correlations still produces precisely the same factor-measurements. These factor-measurements exhibit a succession of dichotomous classifications, sub-classifications, and cross-classifications according to a hierarchical scheme, which is described by the sign pattern of the saturations and measurements. For practical work involving the analysis of empirical answer patterns certain abridged methods are available.

REFERENCES

1. Tchebycheff, P. L. (1859). 'Sur l'interpolation.' Reprinted in *Œuvres* (1907), I, 478-98, 541-60. St.-Petersbourg.
2. Pearson, K. (1901). 'On lines and planes of closest fit to a system of points in space.' *Phil. Mag.*, II, 6th Ser., 559-72.
3. Yule, G. U. (1910). *An Introduction to the Theory of Statistics*. London: Griffin.
4. Burt, C. (1921: 2nd ed. 1947). *Mental and Scholastic Tests*. London: Staples Press.
5. Allan, F. E. (1930). 'The general form of the orthogonal polynomials for simple series with proofs of their simple properties.' *Proc. Roy. Soc. Edin.*, L, 310-23.
6. Walker, D. A. (1931, 1936, 1940). 'Answer pattern and score scatter in tests and examinations.' *Brit. J. Psych.*, XXII, 73-86, XXVI, 301-8, XXX, 248-60.
7. Greenleaf, H. E. H. (1932). 'Curve approximation by functions analogous to the Hermite polynomials.' *Ann. Math. Stats.*, III, 204-17.
8. Aitken, A. C. (1933). 'On fitting polynomials to weighted data by least squares.' *Proc. Roy. Soc. Edin.*, LIV, 1-17.
9. Fisher, R. A. (1934). *Statistical Methods for Research Workers* (5th ed.). Edinburgh: Oliver & Boyd.
10. Burt, C. (1940). *The Factors of the Mind*. London: University of London Press.
11. Ferguson, G. A. (1941). *The Reliability of Mental Tests*. London: University of London Press.
12. Ferguson, G. A. (1941). 'The factorial interpretation of test difficulty.' *Psychometrika*, VI, 323-9.
13. Burt, C., and John E. (1942). 'A factor analysis of the Terman-Binet tests.' *Brit. J. Educ. Psychol.*, XII, 117-21, 156-61.
14. Kendall, M. G. (1946). *Advanced Theory of Statistics*. London: Griffin.
15. Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton: University Press.
16. Fisher, R. A., and Yates, F. (1948). *Statistical Tables*. Edinburgh: Oliver and Boyd.
17. Burt, C. (1950). 'The factorial analysis of qualitative data.' *Brit. J. Psych., Stat. Sect.*, III, 166-85.
18. Stouffer, S. A., et al. (1950). *Measurement and Prediction*. Princeton: University Press.
19. Burt, C. (1951). 'Test construction and the scaling of items.' *Brit. J. Psych., Stat. Sect.*, IV, 95-129.

APPENDIX

Let n be the number of items in the scale, and N the number of persons tested. According to Dr. Guttman, if the scale is perfect, the number of categories of persons between which it can discriminate will be $n + 1$. Let us therefore assume $N = n + 1$. Let j or k ($j, k = 1, 2, \dots, n$) be used to indicate the code-number assigned to the several items when arranged in order of difficulty, and a ($a = n, n - 1, \dots, 1, 0$) be the code-number assigned to the persons when arranged in order of ability: so that n denotes the easiest test and the ablest person. Finally, let x_{ak} stand for the mark obtained by the a th person in the k th item.

In a perfect scale, after the marks have been reduced to deviation form, there will be, for the k th item, k positive marks of $(n - k + 1)$, and $(n - k + 1)$ marks of $-k$. For each item the mean will therefore be $m_k = 0$; the square-sum (q_k say) will be $(n + 1)(n - k + 1)k$, and the variance (σ_k^2) will be $(n - k + 1)k$. Using the familiar formulæ for the sum of the first n numbers and for the sum of their squares, it is then easy to show that the sum of the n variances, $\sum \sigma_k^2$, will be $\frac{1}{6}n(n + 1)(n + 2)$, and that the total mark for the a th person will be $t_a = \sum x_{ak} = \frac{1}{2}(n + 1)(2a - n)$, and his mean mark therefore $\frac{1}{2n}(n + 1)(2a - n)$. The mean marks for the several categories of persons will consequently constitute a perfect linear scale diminishing by regular steps of $\frac{n + 1}{n}$; and their variance, σ_m^2 , will be $\frac{(n + 2)(n + 1)^2}{12n}$.

Let us now form a product-sum matrix by cross-multiplying the marks for every possible pair of items. The result will be, not a synclinal or hierarchical matrix, but what I have called an isoclinal¹ matrix. If we add each column of this matrix, and call the total for the k th item $\sum_j p_{jk}$,

we shall have² $\sum_j p_{jk} = -(n + 1) \sum_{j=0}^{n-k} t_j = \frac{1}{2}(n + 1)^2 k(n - k + 1) = \frac{1}{2}(n + 1)q_k = \frac{1}{2}(n + 1)^2 \sigma_k^2$.

For the correlations we have, by the usual product-moment formula, $r_{jk} = \frac{p_{jk}}{(n + 1)\sigma_j\sigma_k}$.

If we proceed to factorize the matrix of correlations by the method of weighted summation, then (as may readily be shown) the saturations for the first factor will be

$$f_k = \frac{1}{2}(n + 1) \frac{\sigma_k}{n\sigma_m} \quad (i)$$

We can verify this by applying the basic formula $f'_1 R = v_1 f'_1$ (where v_1 denotes the factor-variance and f'_1 the row-vector of factor-saturations used to weight the columns of the matrix). The factor-variance, v_1 , will be given by the sum of the squares of the n saturations, $\sum f_k^2 = \frac{(n + 1)^2 \sum \sigma_k^2}{4n^2 \sigma_m^2}$. On

substituting the values already reached for σ_k and σ_m , this expression reduces to $\frac{1}{2}(n + 1)$. Now, when we premultiply the j th element in the k th column of this matrix by the appropriate saturation

of j , we obtain $f_j r_{jk} = \frac{1}{2}(n + 1) \frac{\sigma_j}{n\sigma_m} \cdot \frac{p_{jk}}{(n + 1)\sigma_j\sigma_k} = \frac{p_{jk}}{2\sigma_m\sigma_k}$. Summing all such values in the k th column, we have

$$\begin{aligned} \sum_j f_j r_{jk} &= \frac{1}{2n\sigma_m\sigma_k} \sum_j p_{jk} \\ &= \frac{1}{2n\sigma_m\sigma_k} \cdot \frac{1}{2}(n + 1)^2 \sigma_k^2 \\ &= \frac{1}{2}(n + 1) \cdot \left\{ \frac{1}{2}(n + 1) \frac{\sigma_k}{n\sigma_m} \right\} = v f_k. \end{aligned} \quad (ii)$$

Hence the values calculated for the saturations by eqn. (i) conform to the requirements of the saturations for the principal axes as stated by Pearson.

¹ See this *Journal*, IV, 1951, p. 47, and refs.

² The minus sign is due to the fact that, with Dr. Guttman's notation, the persons are numbered backwards.



The regressions (w_k), as usual, are obtained by dividing the corresponding saturations by σ_m . We thus have $w_k = \frac{1}{n} \cdot \frac{\sigma_k}{\sigma_m}$. The factor-measurement for the a th person will accordingly be obtained by taking a weighted sum of his standardized marks, viz.,

$$\begin{aligned} \sum w_k \frac{x_{ak}}{\sqrt{q_k}} &= \sum \left\{ \frac{\sigma_k}{n\sigma_m} \cdot \frac{x_{ak}}{\sigma_k \sqrt{(n+1)}} \right\} = \frac{\sum x_{ak}}{n\sqrt{(n+1)}\sigma_m} \\ &= \sqrt{\left\{ \frac{3}{n(n+1)(n+2)} \right\}} \cdot (2a - n). \end{aligned} \quad (\text{iii})$$

The sum of the squares of $(2a - n)$ for $a = 0, 1, 2, \dots, n$ will be $4\sum a^2 - 4n\sum a + (n+1)n^2 = \frac{1}{3}n(n+1)(n+2)$. Hence the effect of the expression under the radical is simply to reduce the values of $(2a - n)$ to standard measure; and these values will evidently form a perfect arithmetical progression decreasing by a constant difference of 2.

It follows that a factorial analysis applied to the standardized marks is fully capable of demonstrating whether or not the marks obtained by the several items form a perfect scale in Dr. Guttman's sense. To measure the degree to which any empirical set of marks approximates to a perfect scale, we could, if we wished, correlate the factor-measurements with the ranks of the several persons (or categories of persons). If we are not interested in determining the rest of the factors, it would for most purposes be sufficient to correlate the unweighted totals, without performing a complete factorial analysis to obtain the most appropriate least-squares weights.

Algebraic expressions for the remaining factors can be reached along similar lines. The order of the curve required to give a perfect fit is the same as the number of degrees of freedom available for the differences between the observed values, namely, $N - 1 = n$. This will therefore be the order of the polynomial expression for the n th bipolar factor, i.e., for the last.

Let us reduce the rank order for the persons to deviation form, and write $x_{0i} = 1$ ($i = 1, 2, \dots, N$) for the general factor-measurements (which are the same for all persons), and $x_{1i} = a_i - \frac{1}{2}(N-1)$ for the first factor-measurements. Then the formulæ for all the factor-measurements can be expressed algebraically as follows.¹ I add in each case the numerical equations so given for the data examined in the preceding pages, where $N = 6$ (cf. Table V).

$$\begin{aligned} \text{G. } \frac{x_0}{\sigma_0} &= \frac{1}{\sigma_0} = \frac{1}{\sqrt{6}}. \\ \text{I. } \frac{x_1}{\sigma_1} &= \frac{1}{\sigma_1} x_1 = \frac{2}{\sqrt{70}} x_1. \\ \text{II. } \frac{x_2}{\sigma_2} &= \frac{1}{\sigma_2} \left\{ -\frac{1}{12} (N^2 - 1) + x_1^2 \right\} = \frac{\sqrt{3}}{4\sqrt{7}} \left\{ -\frac{35}{12} + x_1^2 \right\}. \\ \text{III. } \frac{x_3}{\sigma_3} &= \frac{1}{\sigma_3} \left\{ -\frac{1}{20} (3N^2 - 7) x_1 + x_1^3 \right\} = \frac{\sqrt{5}}{18} \left\{ -\frac{101}{20} x_1 + x_1^3 \right\}. \\ \text{IV. } \frac{x_4}{\sigma_4} &= \frac{1}{\sigma_4} \left\{ \frac{3}{560} (N^2 - 1) (N^2 - 9) - \frac{1}{14} (3N^2 - 13) x_1^2 + x_1^4 \right\} \\ &= \frac{\sqrt{7}}{24} \left\{ \frac{81}{16} - \frac{95}{14} x_1^2 + x_1^4 \right\}. \\ \text{V. } \frac{x_5}{\sigma_5} &= \frac{1}{\sigma_5} \left\{ \frac{1}{1008} (15N^4 - 230N^2 + 407) x_1 - \frac{5}{18} (N^2 - 7) x_1^3 + x_1^5 \right\} \\ &= \frac{\sqrt{7}}{20} \left\{ \frac{11567}{1008} x_1 - \frac{145}{18} x_1^3 + x_1^5 \right\}. \end{aligned}$$

The values for $\sigma_h = \sqrt{\sum_a x_a^2}$ ($a = 1, 2, \dots, N$), the constant used to reduce the values to unitary standard measure, can be obtained by the following equation:

$$\sum x_h^2 = \frac{(h!)^4}{(2h)!(2h+1)!} N(N^2 - 1)(N^2 - 2^2) \dots (N^2 - h^2). \quad (\text{iv})$$

It will be seen that, although the five sets of factor-measurements are uncorrelated, they can nevertheless all be expressed as polynomial functions of one factor only, namely, the first. The order of the polynomial increases step by step; the even factors contain only even powers, while the odd contain

¹ The unstandardized values, denoted by x_h and given within the brackets, are proportional to, but not identical with, those given by Dr. Guttman, which give proportionate values in terms of the smallest integral numbers, like those tabulated by Fisher and Yates ((16), pp. 70f.). To obtain their tabulated figures, the above values for x_h must be multiplied by the fraction at the foot of their tables.

only odd powers: (hence the alternation between symmetry and reversal in the signs of the successive rows of Table V). For practical work the investigator may either take the values from the tables of orthogonal polynomials, or, if these are not available, build them up by means of the following relation¹:

$$x_{h+2} = x_1 x_{h+1} - b x_h, \text{ where } b = \frac{\sum x_{h+1}^2}{\sum x_h^2}.$$

For any factor the saturation for the j th item (f_j) can be obtained from the factor-measurements (p_i), by a process of cumulative summation, using the equation $f_j = \sqrt{\frac{(n+1)}{j(n-j+1)}} \times \sum p_i$, the summation extending to the first j measurements: e.g., for factor BI, taking the figures from line 2 of Table V, $f_2 = \sqrt{\frac{6}{2 \times 4}} \times (.598 + .359) = .866 \times .957 = .8281$, as given in line 2 of Table IXA.

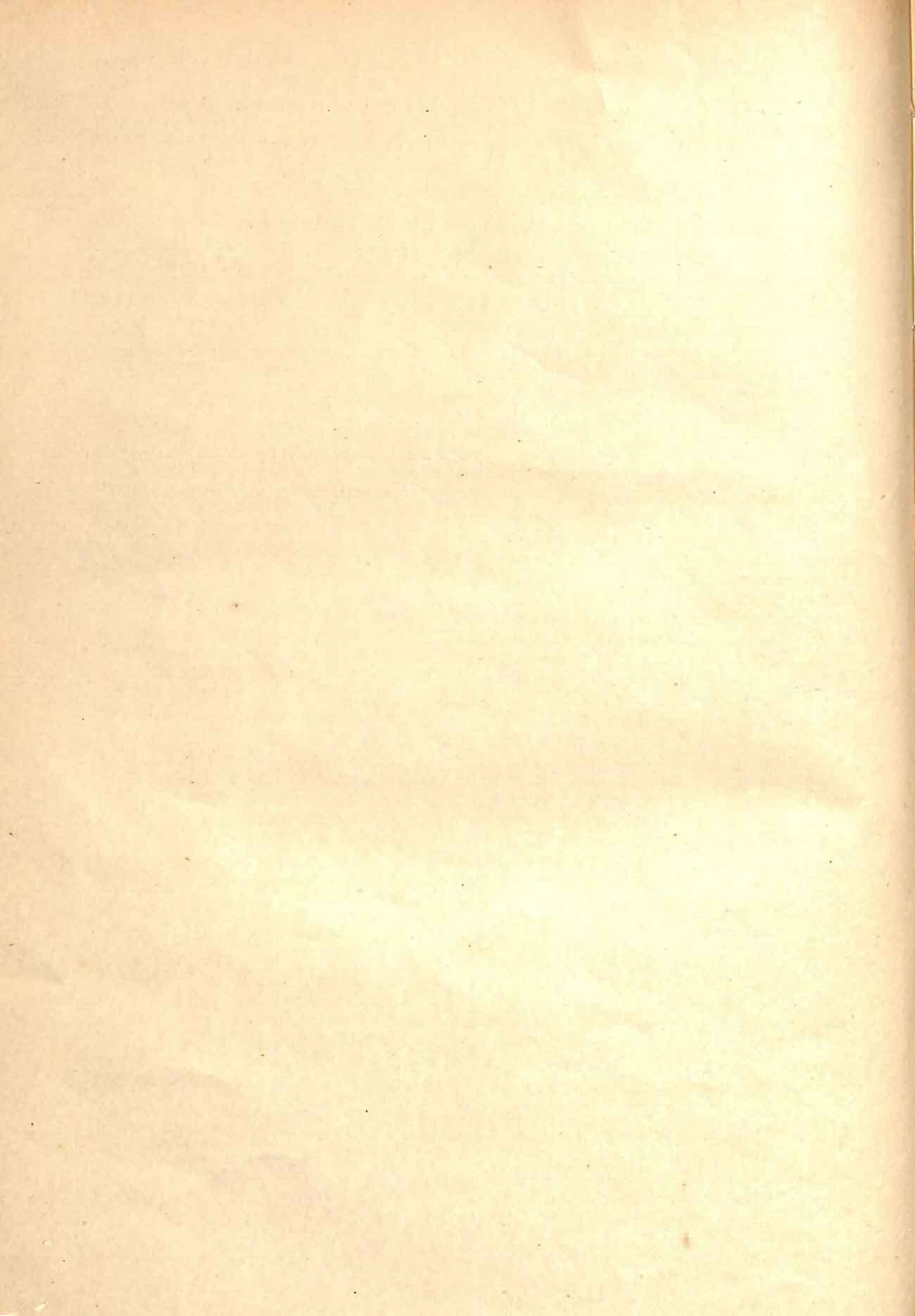
The factor-variances are given by the equations:

$$v_1 = \frac{n+1}{1 \cdot 2}, v_2 = \frac{n+1}{2 \cdot 3}, \dots, v_r = \frac{n+1}{r(r+1)}, \dots, v_n = \frac{n+1}{n(n+1)} = \frac{1}{n}.$$

Their sum is

$$\begin{aligned} \sum v_h &= (n+1) \left\{ \frac{1}{1 \cdot 2} + \frac{1}{2 \cdot 3} + \dots + \frac{1}{n(n+1)} \right\} \\ &= (n+1) \left\{ \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \dots + \left(\frac{1}{n} - \frac{1}{n+1}\right) \right\} = (n+1) \left(1 - \frac{1}{n+1}\right) = n. \end{aligned}$$

¹ The relation is due to Tchebycheff (1). Cf. also (5) and Isserlis, L., 'A Note on Tchebycheff's Interpolation Formula,' *Biometrika*, XIX, 1927, pp. 87f.



COVARIANCE ANALYSIS AND ITS APPLICATIONS IN PSYCHOLOGICAL RESEARCH

By NEIL GOURLAY

Department of Education, University of Birmingham

I. *Problem.* II. *Main Uses of Covariance Analysis.* III. *Comparison of the Precision of Covariance with that of other Procedures.* IV. *Summary and Conclusions.*

I. PROBLEM

Fisher's analysis of variance is now a commonplace in the field of educational and psychological research: its principles and methods of application are expounded in all modern statistical text-books. But, when we turn to its correlate, the analysis of covariance, it is probably true to say that very few text-books provide an adequate treatment of the subject, although some give an excellent account of a particular aspect, and that none gives a complete and critical discussion of its uses in psychology. The wide variation in approach (compare, for example, Lindquist (1), Snedecor (2), McNemar (3), and Kendall (4)) is itself a source of confusion to all but systematic students of statistical method.

This paper therefore attempts to review the main uses of the analysis of covariance with special reference to its applications in psychological research. Its primary object will be to show that, although the statistical analysis remains more or less the same in all cases, the interpretations of the results must vary greatly with the conditions under which the data were obtained. Further, in indicating the correct interpretations for different experimental situations, the opportunity will be taken of pointing out some of the wrong interpretations made in applying the technique. A comparison will also be made of the precision of covariance with that of other methods; and it will be shown that there are frequent occasions when a simpler technique can be employed with little loss in precision and a considerable reduction in the amount of computational labour. Bartlett (5)¹ has published an excellent but brief paper on covariance analysis. But this appeared in a journal which is not readily accessible to the average educational and psychological investigator, and was written mainly for the agricultural experimenter; moreover, the misinterpretations of which Bartlett speaks are still to be found.

II. MAIN USES OF COVARIANCE ANALYSIS

Whatever other interpretations may be given, an analysis of covariance can always be regarded as a means of describing the manner in which any particular variate varies with respect to other variates within a given number of groups or populations. Covariance is therefore always applicable when samples from the same or different populations are measured for two or more variates, whether or not the measurements are made at the same time. However, every hypothesis put for-

¹ I have to acknowledge that, while most of my ideas on the analysis of covariance had taken shape before I read Bartlett's paper, their final form owes much to it. I am also indebted to him for certain comments he has been good enough to make on my manuscript and take this opportunity of expressing my thanks.

ward for testing requires that some set of statistical conditions shall be satisfied before the technique can be validly applied. These differ with the different types of hypothesis, and are not always clearly defined in the text-books. They will be stated as explicitly as possible in the course of the following discussion.

In this section the various applications will be considered with reference to three types of case :

- (a) where covariance is used to increase the precision of an experiment ;
- (b) where it is a means of analysing the relation between variates which have been measured after treatment of experimental groups ;
- (c) where it is applied to data obtained from non-random groups.

For simplicity, the discussion will be confined to the case of two variates only, i.e., we shall consider the variation of one variate, the dependent, relative only to one other, the independent, variate. This is by far the commonest situation in educational and psychological research ; and a similar argument holds when more than one independent variate is involved.

(a) *The Analysis of Covariance as a Means of Increasing the Precision of an Experiment.* This is probably the most important function that the analysis of covariance can fulfil in educational and psychological research. It is the only application considered by Fisher and Lindquist in their text-books ((6), (7), and (1)) : and it is one of the two discussed by Edwards (8). On the other hand, it receives scant treatment from Snedecor (2), and is not treated at all in many other text-books, e.g., Johnson's (9). Johnson admittedly, in one of his examples, speaks of covariance as 'increasing the precision of the experimental comparisons' ; but, as I shall point out later, the phrase is not legitimate in such a context.

When one speaks of increasing the precision of an experiment by the use of covariance analysis, it is evidently implied that this increase is relative to the precision obtained by other methods of analysis. Here we shall compare covariance with the simple analysis of variance only. For ease of exposition, let us make the comparison in terms of a simple experiment designed to compare the differential effects of, say, two or more methods of teaching a topic in some school subject. For both techniques the experimental procedure is the same except in one respect : in particular it is essential that the experimental groups shall form *random* samples from the given population. But, when using covariance, besides applying an appropriate criterion as a test at the conclusion of the experiment, it is also necessary that the groups shall be tested prior to the experiment and that the test adopted shall correlate well with the criterion. Often the same test is used, thus ensuring the maximum correlation. In the covariance analysis the scores in the initial and final testing become the values for the independent and dependent variate respectively. The two variates may be conveniently referred to as the *initial* and *final* variates.

The null hypothesis is the *same* for both analyses : otherwise it would be meaningless to talk of greater precision in the case of covariance. A simple statement of the null hypothesis is that no significant difference exists between the methods of teaching. Alternatively, we may say that the scores for the experimental groups in the criterion-test are random samples from the same population ; or, less generally, that the population means for the two groups working under different teaching methods are equal, i.e., $\mu_{py} = \mu_{qy}$, where μ_{py} and μ_{qy} are the population means for the *p*th and *q*th methods groups in the criterion-test (*y*).

An analysis of variance tests the hypothesis by providing two independent estimates of the common population variance—the 'between methods' and 'within methods' variances based on $(S - 1)$ and $(N - S)$ degrees of freedom respectively, where *S* is the number of methods and *N* is the total number of subjects employed in the experiment. The variances are tested for heterogeneity by means of Snedecor's *F*-test or Fisher's *z*-test. Necessary assumptions in making the test are that the scores 'within methods' are normally (or approximately normally) distributed and that the methods do not give rise to significant differences in the variances for the different experimental groups (the condition of homogeneity of variance).¹

¹ Although in this paper I have some criticism to make of Johnson's treatment of the analysis of covariance (9), his treatment of the analysis of variance is to be recommended in that he always tests for homogeneity of variance—a precaution that many investigators often overlook. But, rather inconsistently, he does not consider it necessary to test the underlying assumptions required for an analysis of covariance.

An analysis of covariance, on the other hand, utilizing the fact that the differences between final scores are to some extent a reflection of differences between the initial scores, removes the variance due to these initial differences from the final variances. It thus reduces the size of the error variance, and so increases the precision of the experiment. The algebraic details are given in the text-books (e.g., (1), pp. 191-3). Briefly the steps are:

1. An analysis of the sums of squares for the initial and final scores (i.e., the independent and the dependent variate respectively).
2. An analysis of the sum of products for the two variates.
3. Calculation of the residual sum of squares for the dependent variate (*a*) when the *total* sums of squares and products are used to fit a regression line to the total data and (*b*) when the *within methods* scores of squares and products are used to fit a regression line of common slope to each of the methods groups.
4. The subtraction of the residual sum of squares for *within methods* from the *total* residual sum of squares to give the *between methods* residual sum of squares.

One degree of freedom is lost in fitting the regression line; and the results of the analysis can therefore be tabulated as follows:

Residual sums of squares	d.f.
Between methods	$S - 1$
Within methods	$N - S - 1$
Total	$N - 2$

On the null hypothesis the residual sums of squares *between methods* and *within methods* provide independent estimates of the same variance, namely, the residual variance in the parent population. Hence the hypothesis is still tested by the *F*-test. The underlying assumptions are clearly stated by Lindquist ((1), pp. 195, 196):

- (i) The methods groups must be *random* samples from the *same* population;
- (ii) The regression of final on initial scores must be linear;
- (iii) The regression coefficients must not differ significantly from group to group;
- (iv) Within any group the dependent variate must be distributed normally for any given value of the independent variate, and have a constant variance, the same for all groups, namely, the residual variance already mentioned.

For each of these assumptions tests are available; and, strictly speaking, they should form part of any analysis of covariance. The test for condition (iii) is given by Kendall (4), Snedecor (2), and Jackson (10). The test for homogeneity of residual variance is similar to that used in the analysis of variance: an example of its application in covariance analysis will be found in Jackson ((10), pp. 75-7). No assumption need be made about the distribution of the independent variate; and consequently the condition of normality is not necessary (condition (iv) applies only to the partial variate).

What is the actual gain in precision achieved by covariance as compared with a simple analysis of variance? Since precision is measured by the reciprocal of the error variance ((7), p. 182), the question can be answered by comparing the error variances. The ratio of the two precisions proves to be very nearly $(1 - r^2)$, where r is the product-moment correlation within groups. This follows from the well-known fact that the residual sum of squares within groups for the dependent variate is equal to $\Sigma y^2(1 - r^2)$, where Σy^2 is the sum of squares within groups for that variate. Remembering that the residual sum of squares 'within groups' it will be seen that the exact value of the precision ratio is $(N - S)(1 - r^2)/(N - S - 1)$, i.e., slightly more than $(1 - r^2)$.

In more complex experiments, where one or more factors are controlled and not randomized, the error variance may be an interaction term; but the above result still holds, except that r is now derived from the sums of squares and products for the interaction term. It is to be noted, however, that, when the interaction term is based on a small number of degrees of freedom, the one degree that is lost in analysis of covariance may considerably reduce the gain in precision.

(b) *The Use of Covariance to Study the Relation between Variates Measured after Treatment of Experimental Groups.* This is the application with which Snedecor (2) is mainly concerned; it is briefly considered by Edwards (8).

Here we are again dealing with an experimental situation, i.e., one where principles of randomization, statistical control, etc., are deliberately introduced. But now both variates are measured after the application of the experimental treatments: that is, both may be affected by the treatments, and thus show significant differences between groups.

Since both are now 'final' variates, it would appear arbitrary which we choose as the independent variate; but in practice the choice is determined by the question which has prompted the analysis. Four possibilities arise¹ :—

- (i) We may know on *a priori* grounds that there are no differences between treatments for the independent variate;
- (ii) Significant differences between groups are found for the independent variate²;
- (iii) We may know *a priori* that there are real differences between treatments for the independent variate, although a test of group differences in this variate does not prove significant;
- (iv) We may have no *a priori* knowledge about the differences between treatments for the independent variate; but, on testing them, the differences are found to be non-significant.

In case (i) the situation is the same as that discussed in subsection (a): and in applying the analysis of covariance we are (as before) obtaining a more precise test. To illustrate from the educational field, one could carry out an experiment on method, e.g., in the teaching of Arithmetic, and at the end of the experiment give tests in both Arithmetic and Intelligence to the groups. Then, if one could say *a priori* that there was no difference between the groups in intelligence (and the data supported this), the scores for the Intelligence test could be used to improve precision in testing the differences between groups in the Arithmetic test. This is not the best procedure. As already stated, maximum precision is obtained by testing the groups initially with a test which has a high correlation with the final test. However, a situation might arise in which an investigator through ignorance failed to set an initial test, and a statistician called in later tried to remedy the position by setting a test at the end which correlated well with the final test, and for which there was known to be no difference between groups.

Turning to case (ii), namely, that in which there are significant differences between groups in the independent variate, we shall assume that there are also significant group differences in the dependent variate: otherwise there would be little point in using covariance. That is the situation always assumed in text-books concerned with this type of application. Since in this case the statistical details are straightforward and trouble arises only when less technical descriptions are attempted, it will be advantageous to consider the former first. The algebraic procedure is the same as for subsection (a); and the same variances are involved in the *F*-test. The necessary assumptions (p. 27) are also the same. But the hypothesis tested is different.

We can say that what we are testing is the hypothesis that the regression line (for the dependent variate on the independent variate) is the same line (though not necessarily of the same gradient) for all treatments; or that for a given value of the independent variate the estimated value of the dependent variate is the same for all treatments; or again that the means of the bivariate populations corresponding to the different groups are collinear, the line of collinearity being the common regression line. Algebraically, if μ_{px} , μ_{py} and μ_{qx} , μ_{qy} are the population means for any two treatments *p* and *q* (*x* and *y* representing the independent and dependent variates respectively), the hypothesis is that $(\mu_{py} - \mu_{qy})/(\mu_{px} - \mu_{qx}) = \beta$, where β is the coefficient of regression in the common population.

It must be emphasized that this is not the same as the hypothesis involved in subsection (a), where $\mu_{px} = \mu_{qx}$ by virtue of the design of the experiment, and the hypothesis to be tested is simply $\mu_{py} = \mu_{qy}$ —the same hypothesis as is tested by the simple analysis of variance.³ It follows from the definition of the term 'precision' that this application of covariance cannot be regarded as a case in which the purpose is to increase precision. If the expression 'increase of precision' is to have any meaning, the hypothesis tested must be the same when the precision technique is used as when it is not.⁴

¹ As in subsection (a), it will be convenient to speak in terms of a simple experiment on methods of treatment.

² If an analysis of variance is used to test differences between groups in the independent variate, the minimum assumptions for an analysis of covariance (p. 27) must be supplemented by the assumption of normality (or approximate normality) within groups for the independent variate. Both variates must then be normally distributed.

³ Mathematically the hypothesis of subsection (a) may be considered as a limiting case of the other.

⁴ I therefore disagree with Snedecor ((2), p. 323) when he regards 'increase of precision' as one of the features of this application. The remarkable thing is that later (pp. 334, 335) he appears to be quite clear about the distinction.

Let us now attempt a less technical description of the hypothesis stated above. Still keeping to statistical terms, we can say that we are seeking to discover whether the relation which the dependent variate bears to the independent variate is the same *between groups* as *within groups*; more loosely, whether the 'between group' differences in the dependent variate are "adequately accounted for" (as the text-books sometimes put it) by the 'between group' differences in the independent variate. I give two examples.

(1) Two groups of young children, backward mentally and educationally, are assigned at random to two different kinds of environment. Some years later they are tested for intelligence and educational attainment. To determine whether a difference between the groups in attainment is accounted for by their difference in intelligence, an analysis of covariance might be suitably used.

(2) In an experiment to compare methods of teaching composition, the experimental groups may be tested at the end of the experiment for ability in composition and in the mechanical aspects of English. Once again we might use covariance to discover whether group differences in the second variate 'explain' the group differences in the first.

There is yet another popular way of expressing the purpose in this case. We can speak of testing for differences between groups in the dependent variate after the latter have been 'corrected' for differences between groups in the independent variate, or, as Snedecor put it ((2), p. 320), after 'due allowance is made for' such differences. So long as such expressions are understood to mean no more than is contained in the previous statistical statements, they are permissible. There is one interpretation which must not be made. It must not be thought that, when covariance is applied to make "due allowance for" the independent variate, we are then testing the same hypothesis as would be involved in an experiment where it was arranged *experimentally* that there was no difference between groups in that variate.¹

Case (iii) is easily dismissed. The interpretation of an analysis of covariance for this situation is the same as for case (ii).

Case (iv) might at first sight be put with case (i).² But, if one considers the case in which normally distributed groups show no significant differences for either variate but a significant overall difference when a Λ -test is applied,³ it will be seen at once that the identification would be wrong. The significant overall difference must be due to real differences between groups in either the one variate or the other or in both; and the data do not allow us to accept one of these alternatives and reject the others. It follows that we cannot speak of using the one variate as the independent variate of a covariance analysis in order to increase the precision in testing differences as regards the other variate. Hence case (iv) cannot be identified with case (i), but should rather be interpreted along the lines adopted for case (ii).

(c) *The Analysis of Covariance Applied to Non-Random Groups.* In the previous applications the groups were random samples from the same population. In a third type of application they may be non-random or (to use Lindquist's expression) 'intact.' Instances from educational enquiries are the pupils in a school or grade or class. The variates may be measured at the same time or at different times (cf. Kendall (4), Johnson (9), McNemar (3), and Jackson (10)).

As before, we will consider the purely statistical aspect first. The algebraic procedure is the same: in this respect there are no differences between the three types. Further, the underlying assumptions are the same, except that the condition of randomness no longer holds. It must be emphasized that, with non-random groups, heterogeneity is more likely than in the previous cases. The researches quoted by Lindquist ((1), pp. 134-41, 223-7) suggest that, in dealing with intact school groups, heterogeneity of variance and correlation is the rule rather than the exception. The same is probably true of regression. Hence it is all the more urgent that the underlying assumptions of homogeneity of regression and residual variance should first be tested.

What statistical hypothesis is involved in this type? From subsections (a) and (b), the reader will see that the only valid hypothesis is that already stated for subsection (b), case (ii). If this is borne in mind, there will be no difficulty in examining the validity of

¹ An illustration of this point is a simple experiment on methods of instruction where the experimental groups do not receive the same total amount of instruction time. An analysis of covariance which allows for the differences in times will not necessarily give the same result (on applying the F -test) as an experiment where the times were made the same. The reader should be able to give the reason why this is so.

² Bartlett (5) does not appear to differentiate between cases (i) and (iv).

³ For an example of the use of the Λ -test, see this *Journal* (11).

less technical interpretations. Consider first whether covariance can be used to compare treatments, methods, etc., when applied to non-random groups instead of random samples. An educational investigator sometimes supposes that, if he uses initial scores (or intelligence-test scores) as the independent variate, he can validly apply covariance to non-random groups.¹ But, although the analysis makes an allowance for differences in initial scores or in intelligence-test scores (or in both), there is no guarantee that there may not be other differences between the groups operating to produce final differences which will be unattributable to the treatments. Unless other evidence is available, the danger of drawing a false conclusion is equally great when there are no differences between groups in the initial or final tests. Kendall's example ((4), pp. 240-2) illustrates this point. In his example the groups show no significant differences in the initial test but significant differences in the final test *after being subjected to the same treatment*.²

It follows that this third type of application, like case (ii), cannot be regarded as a method of increasing precision. I would therefore disagree with Johnson ((9), pp. 310-24) when he describes his use of covariance as 'increasing the precision of the experimental comparisons.' Obviously, since the same statistical hypothesis is involved as in subsection (b), case (ii), the interpretations which were suggested there can also be made for the present case. There is no need to repeat them. This does not exhaust the possibilities of this particular type. Kendall's example shows the value of covariance in prediction problems. In that example "a number of recruits are given a preliminary test to ascertain their suitability for a certain course of training; at the end of the training they undergo a proficiency test." One of the issues which the analysis is intended to decide is "whether the same rejection standards in the preliminary tests can be applied to all recruits, irrespective of their town of origin." The reader will note that, whatever specific purpose covariance is made to serve, the general purpose is always that of looking for homogeneity or heterogeneity—as in any analysis of variance. Jackson ((10), p. 74) expresses the same idea when he says that covariance decides whether the groups can be combined for the tests involved.³

III. COMPARISON OF THE PRECISION OF COVARIANCE WITH THAT OF OTHER PROCEDURES

In the preceding section I have endeavoured to indicate in what cases an analysis of covariance may be validly said to yield increased precision, and how the precision of covariance compares with that of a simple analysis of variance. I now propose to compare it with two other statistical procedures.

(a) *Matching Techniques*. Consider again a simple experiment on methods. It is well known that precision can be improved by matching the pupils, prior to the experiment, in one or more tests—in pairs if only two methods are to be compared, in threes if (as is somewhat rare) three methods are involved, and so on. The pupils in each matched group

¹ The first sentence of the second paragraph of McNemar's chapter on covariance (3) suggests that he takes this view.

² Kendall states (p. 243) that, by using the analysis of covariance to allow for initial differences, we obtain the effects of training alone on the groups. I would suggest that the residual differences between groups will be due partly to the effects of training on the groups, but partly to any difference in content between the preliminary and final tests that does not result from the training.

³ An unusual feature of Jackson's analysis ((10), p. 80) is that, in testing for differences between classes in the second test, after correcting for differences found in the first, he uses, not the variance derived from the residual sum of squares between classes (based on $(s-1)$ d.f.), but the variance derived from the sum of squares due to departure from linear regression between classes (based on $(s-2)$ d.f.). In other words, in correcting final differences for initial differences he is using not the regression *within classes* but the regression *between classes*. A valid interpretation can be given for Jackson's procedure (see Welch (12)); but I do not know whether he had this in mind.

are then assigned at random to one or other of the methods. For the final scores the analysis of variance will be tabulated as follows :

Sums of Squares				d.f.
Between methods	..	..	..	$p - 1$
Between matched groups	..	..	..	$n - 1$
Error	..	..	..	$(n - 1)(p - 1)$
Total				$np - 1$

where p is the number of methods and n the number of matched groups.

If the matching were perfect the error variance, it will be seen, would here furnish an estimate of the same variance as is estimated by the error variance in the analysis of covariance for the initial and final scores—namely, the residual variance within groups for the total population. As a rule matching is far from perfect, particularly if more than two methods are involved and the number of pupils is limited. Hence, on this count alone, the matching method never achieves as high a precision as the analysis of covariance. The precision of the matching procedure is also reduced owing to the fact that $(n - 1)$ d.f. are lost in 'taking out' the variance between matched groups. The loss amounts to 50 per cent. when only two methods are involved, 33 per cent. when three are involved, and so on. In some cases it may even render the method less precise than a simple analysis of variance. For this reason and because of the practical difficulty of matching pupils, experienced investigators seldom use a matching technique with experiments on methods and the like, if an analysis of covariance can be made.

Similar remarks apply to experiments on factorial design where other classifications besides that of treatments enter into the analysis. Thus, for an experiment on methods, replicated in several schools, the precision might perhaps be improved by matching the methods groups within schools; but the same object can be achieved more easily and effectively by using random groups or even intact classes, and then applying an analysis of covariance.

(b) *Analysis of Variance of Gains (or Differences between Initial and Final Scores)*. In his *Design of Experiments* ((7), pp. 166-9) Fisher discusses certain alternatives to the analysis of covariance that have been used by investigators, namely :

- (i) an analysis of the variance of absolute gains of final over initial measurements ;
- (ii) an analysis of the variance of such gains expressed as a fraction or percentage of the initial measurements.

While recognizing that, in the usual agricultural experiment, such methods may not give results greatly different from an analysis of covariance, he points out that they entail arbitrary corrections for initial differences, and consequently never achieve the precision of a covariance analysis. Further, in the absence of other evidence, the experimenter has no assurance that the corrections are appropriate to his data, nor can he tell how much precision he has lost by employing them.

Fisher does not completely condemn these methods. And accordingly I propose to consider the first of them, and indicate (for experiments in the educational field) what conditions must be satisfied to obtain adequate results, i.e., results which do not fall short of the precision secured by an analysis of covariance.

The usual conditions of normality (or approximate normality) and homogeneity of variance will again hold good; but in this case they need apply only to the gains (or differences). Let us denote the initial and final variates by X and Y , and consider the analysis of the variance of the differences $Y - X$. The hypothesis to be tested is obviously still the same as for an analysis of variance (of final scores) or covariance; it will in fact be that $\mu_y - \mu_x$ (in the usual notation) is the same for all groups; it thus reduces to the hypothesis that μ_y is the same for all groups (since μ_x is the same for all groups as a result of the experimental design).

Let σ_x^2 , σ_y^2 and ρ denote the population values of the variances within groups of the X and Y variates and of the product-moment correlation within groups respectively. Then the precision in the case of the covariance analysis is the reciprocal of the residual variance within groups, which, on the average, has the population value $\sigma_y^2(1 - \rho^2)$. For the precision in the case of the analysis of variance of differences we must consider the variance within groups which will have, on the average, the population value $\sigma_{y-x}^2 = \sigma_y^2 - 2\rho\sigma_x\sigma_y + \sigma_x^2$.

The loss in precision is therefore on the average ¹

$$\frac{1}{\sigma_y^2(1-\rho^2)} - \frac{1}{\sigma_y^2 - 2\rho\sigma_x\sigma_y + \sigma_x^2} = \frac{(\sigma_x - \rho\sigma_y)^2}{\sigma_y^2(1-\rho^2)(\sigma_y^2 - 2\rho\sigma_x\sigma_y + \sigma_x^2)};$$

or, expressed as a fraction of the maximum precision (given by the analysis of covariance),

$$\frac{(\sigma_x - \rho\sigma_y)^2}{\sigma_y^2 - 2\rho\sigma_x\sigma_y + \sigma_x^2} = \frac{\left(\frac{\sigma_x}{\sigma_y} - \rho\right)^2}{\left[\left(\frac{\sigma_x}{\sigma_y}\right)^2 - 2\rho\left(\frac{\sigma_x}{\sigma_y}\right) + 1\right]};$$

a quantity which is a function of ρ and the ratio σ_x/σ_y .

As will be seen, there is little loss of precision when σ_x/σ_y is approximately equal to ρ . Where σ_x/σ_y is not of the same order as ρ , and where a straightforward analysis of variance of differences would involve too great a loss in precision, the analysis of covariance can still be avoided by multiplying the values of the X and Y variates by factors, a and b say (so that $\text{var}[aX] \div \rho^2 \text{var}[bY]$), and then carrying out an analysis of variance of the differences $bY - aX$. This is sometimes a practical procedure.

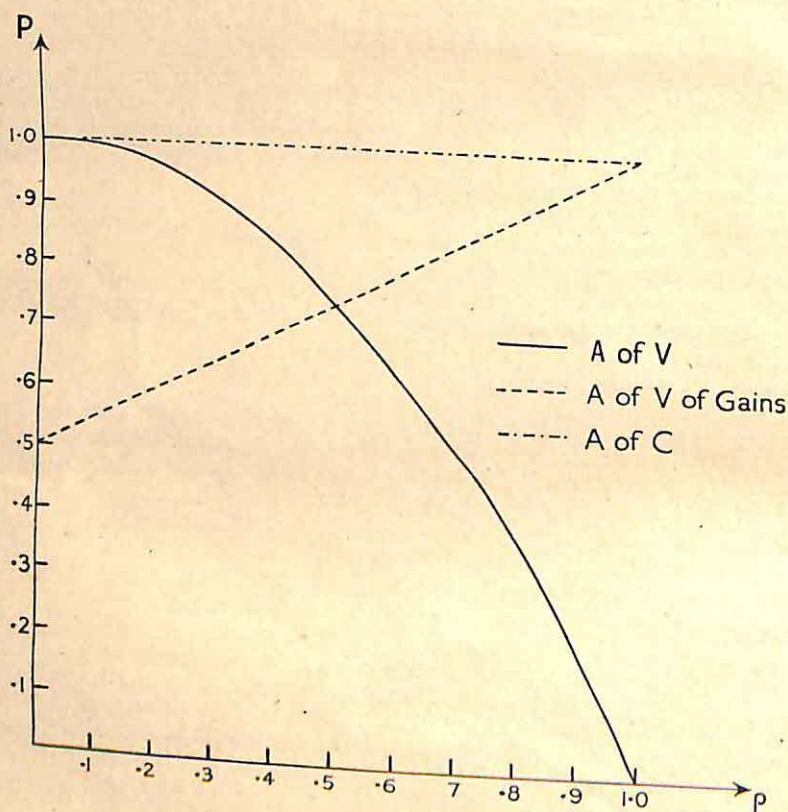


FIG. 1.—Precision of methods of analysis.

The diagram shows the variation of precision (P) with the correlation (ρ) for three methods of treating experimental data. The precision for the analysis of covariance has been taken as unity.

¹ To simplify the argument, I am here ignoring the loss of 1 d.f. incurred by an analysis of covariance.

In educational experimentation a common situation is to have σ_x equal (or approximately equal) to σ_y . This arises (a) when both the initial and final variate values are the scores on a standardized test, and (b) when scores on non-standardized tests have to be normalized before analysis in order that any pronounced skewness in score distribution may be removed and the condition of normality satisfied. Writing $\sigma_x = \sigma_y$ in the above expression, the loss in precision becomes $(1 - \rho)/2$. It will be seen that it is less than 10 per cent. if ρ is greater than .8 (which is often the case if the same test is used both at the beginning and end of the experiment) and only 15 per cent. for ρ equal to .7. Taking the precision of the analysis of covariance as unity, Fig. 1 shows how the precision of the analysis of variance and of an analysis of variance of differences varies with the value of ρ . The same considerations hold good in more elaborate experiments where several factors are controlled, the only difference being that σ_x^2 , σ_y^2 , and ρ now apply to the population values of the variances and correlation for the interaction term providing the estimate of error variance.

The reader may wonder why in such cases one should prefer an analysis of variance of differences to an analysis of covariance. The answer is simply that it saves considerable computational labour, often as much as two-thirds, particularly when the experiment involves several factors. It also avoids the loss of 1 d.f. involved in an analysis of covariance. A good practical procedure is to try the analysis of variance of differences first and to use the analysis of covariance in cases of borderline significance.

IV. SUMMARY AND CONCLUSIONS

1. The various uses of the analysis of covariance have been considered with special reference to its application in psychology and have been discussed under three main headings :

- (a) where covariance is used to increase the precision of an experiment ;
- (b) where it is a means of analysing the relation between variates which have been measured after treatment of experimental groups ;
- (c) where it is applied to data obtained from non-random groups.

2. In each case, special attention has been given to the following points ;

- (i) the necessary statistical conditions for the valid application of covariance analysis ;
- (ii) the exact nature of the statistical hypothesis involved ;
- (iii) the conclusions that can validly be drawn from the results of testing these hypotheses ;
- (iv) misinterpretations of covariance analysis which have been made in text-books and other publications.

3. An attempt has been made to define the particular conditions under which it is correct or incorrect to state that the use of covariance 'increases precision.'

4. The analysis of covariance has been compared in some detail with other methods of analysis—the analysis of variance for random and matched samples and the analysis of variance of gains (or differences). It has been shown that, while the maximum precision is normally obtained by means of covariance, in certain circumstances very little precision will be lost by applying an analysis of variance to the gains (or differences)—a technique which is computationally much simpler.

REFERENCES

1. Lindquist, E. F. (1940). *Statistical Analysis in Educational Research*. New York : Houghton Mifflin Co.
2. Snedecor, G. W. (1948). *Statistical Methods*. Iowa : Iowa State College Press.
3. McNemar, Q. (1949). *Psychological Statistics*. New York : John Wiley & Sons, Inc.
4. Kendall, M. G. (1945). *The Advanced Theory of Statistics, II*. London : Griffin and Co., Ltd.
5. Bartlett, M. S. (1936). 'A note on the analysis of covariance.' *J. Agr. Sci.*, XXVI, 488-91.
6. Fisher, R. A. (1946). *Statistical Methods for Research Workers*. London : Oliver and Boyd.
7. Fisher, R. A. (1949). *The Design of Experiments*. London : Oliver and Boyd.
8. Edwards, A. L. (1950). *Experimental Design in Psychological Research*. New York : Rinehart & Company, Inc.
9. Johnson, P. O. (1949). *Statistical Methods in Research*. New York : Prentice-Hall, Inc.
10. Jackson, R. W. B. (1940). *Application of the Analysis of Variance and Covariance Method to Educational Problems*. Bulletin XI, Dept. of Educ. Research, University of Toronto.
11. Rao, C. R., and Slater, P. (1949). 'Multivariate analysis applied to differences between neurotic groups.' *Brit. J. Psych., Stat. Sect.*, II, 17-29.
12. Welch, B. L. (1935). 'Some problems in the analysis of regression among k samples of two variables.' *Biometrika*, XXVII, 145-60.

THE DISTRIBUTION OF TOTAL RANK VALUE FOR ONE PARTICULAR OBJECT IN m RANKINGS OF n OBJECTS

By J. W. WHITFIELD

Psychological Laboratory, University College, London

I. Problem. II. Principle of Distribution. III. Moments. IV. Example.

I. PROBLEM

It is sometimes necessary to pick out for particular study one object only in a series, as when an experimental object is presented with control objects for preference ranking or some other form of judgement. A similar situation arises in experimental social psychology, when *one* person in a group is instructed to play a predetermined role, and each of the other members of the group is asked to rank his fellow members (including the experimental person) for certain characteristics. In these cases the experimenter is concerned with the ranks given to the specific object or individual. For this Kendall's (1948) treatment of the general m ranking problem is inappropriate ; and the following is suggested.

II. PRINCIPLE OF DISTRIBUTION

Consider one ranking of the n objects. The particular object can be assigned any rank from 1 to n ; and in the null case each value will have an equal probability. Interest lies in whether the object is ranked high or low, according to the experimental hypothesis under investigation. If a second ranking is considered, the distribution of the total of the ranks assigned to the object can take values from 2 to $2n$; but the null case no longer gives equal probability for each total value. The frequencies in the null case are made up in the following way. With a single ranking only ($m = 1$) we have :

	Rank Values										Total
	1	2	3	.	.	.	.	.	($n-1$)	n	
Frequency	1	1	1	.	.	.	.	.	1	1	n

Adding a second ranking to each possible value of the first, we obtain for $m = 2$:

1st Rank Value	Total Rank Values												Total (f)			
	2	3	4	.	.	.	n	(n + 1)	(n + 2)	.	.	.		(2n - 2)	(2n - 1)	(2n)
1	1						1	1								n
2		1	1	.	.	.	1	1	1							n
3			1	.	.	.	1	1		1						n
.				.	.	.	1			.	.	.	1	1		n
n - 1				.	.	.	1	1	1	.	.	.	1		1	n
n								1	1	.	.	.				
Total Frequency	1	2	3	.	.	.	(n - 1)	n	(n - 1)	.	.	.	3	2	1	n ²

This illustrates the general principle that the frequency distribution for m rankings of n objects consists of the distribution for $m-1$ rankings repeated n times in a simple staggered fashion—a principle which simplifies the computation of the exact values and facilitates the general expression for the second and fourth moments. In the appendix (pp. 38-40) tables of the exact probability values are given for m and n up to 8. For higher values of m or n a normal deviate approximation is possible.¹

III. MOMENTS

It will be seen that the distribution is symmetrical, and therefore the higher odd order moments are zero. The mean total rank value is $\frac{1}{2}m(n+1)$, and deviations from this are to be tested. Let x be the deviation.

Variance

$$\Sigma x^2 = n \Sigma_{m-1} x^2 + 2n^{m-1} \left\{ 1^2 + 2^2 + 3^2 + \dots + \left(\frac{n-1}{2} \right)^2 \right\} \dots \dots (n \text{ odd})$$

$$\text{or} = n \Sigma_{m-1} x^2 + 2n^{m-1} \left\{ \left(\frac{1}{2} \right)^2 + \left(\frac{3}{2} \right)^2 + \left(\frac{5}{2} \right)^2 + \dots + \left(\frac{n-1}{2} \right)^2 \right\} \dots (n \text{ even})$$

(since $\Sigma_{m-1} x^2$ is the sum of x^2 in the m ranking case, when all the n^m equiprobable occurrences are considered)

$$= n \Sigma_{m-1} x^2 + \frac{n^m(n^2-1)}{12}.$$

$$\text{Thus } n \Sigma_{m-1} x^2 = n^2 \Sigma_{m-2} x^2 + \frac{n^m(n^2-1)}{12},$$

$$\text{and hence } \Sigma_m x^2 = \frac{mn^m(n^2-1)}{12},$$

and the variance therefore

$$= \frac{m(n^2-1)}{12}.$$

Fourth Moment

$$\Sigma_m x^4 = n \Sigma_{m-1} x^4 + 2 \times 6 \Sigma_{m-1} x^2 \left\{ 1^2 + 2^2 + 3^2 + \dots + \left(\frac{n-1}{2} \right)^2 \right\} + 2n^{m-1} \left\{ 1^4 + 2^4 + 3^4 + \dots + \left(\frac{n-1}{2} \right)^4 \right\} \dots \dots (n \text{ odd}),$$

¹ The results to be expected with perfectly random rankings may be compared with those which we should obtain if each of the m judges were to spin a perfect n -sided teetotum, with its sides numbered 1 to n , and we were then to add up the m scores: for the total scores obtained in this way the theoretical frequency-distribution can be deduced by the aid of the multinomial theorem. (I am indebted to Professor Burt for this observation. He points out that "the frequency for a total rank of x would then be given by the coefficient a^x in the product of the binomial expansions of $(1-a^n)^m(1-a)^{-m}$. With a little simplification the formula for the coefficients reduces to

$$\frac{m(m+1) \dots (x-1)}{(x-m)!} - m \cdot \frac{m(m+1) \dots (x-n-1)}{(x-n-m)!} + \frac{m(m-1)}{2!} \cdot \frac{m(m+1) \dots (x-2n-1)}{(x-2n-m)!} - \dots,$$

where the series is continued so long as none of the factors are negative. For tail-values, only the first one or two terms will require to be calculated. To find the probabilities we divide by n^m . The moments can be obtained in the usual way from the factorial moment generating function.")

This approach yields the same figures as are given in the appendix, and would make it possible to deal with the more general type of problem when equiprobability cannot be assumed.

with corresponding values for n even,

$$= n \sum_{m-1} x^4 + 12 \sum_{m-1} x^2 \frac{n(n^2-1)}{24} + 2n^{m-1} \frac{n(n^2-1)(n^2-4)}{160} + 2n^{m-1} \frac{n(n^2-1)}{96},$$

$$= n \sum_{m-1} x^4 + (m-1)n^{m-1}(n^2-1) \frac{n(n^2-1)}{24} + \frac{n^m(n^2-1)(n^2-4)}{80} + \frac{n^m(n^2-1)}{48}.$$

Therefore $n \sum_{m-1} x^4 = n^2 \sum_{m-2} x^4 + (m-2)n^{m-1}(n^2-1) \frac{n(n^2-1)}{24} + \frac{n^m(n^2-1)(n^2-4)}{80} + \frac{n^m(n^2-1)}{48}.$

Hence $\sum_m x^4 = \frac{n^m(n^2-1)^2}{24} \{ (m-1) + (m-2) + \dots + 1 \} +$

$$\frac{mn^m(n^2-1)(n^2-4)}{80} + \frac{mn^m(n^2-1)}{48}$$

$$= \frac{n^m(n^2-1)^2 m(m-1)}{48} + \frac{mn^m(n^2-1)(n^2-4)}{80} + \frac{mn^m(n^2-1)}{48}.$$

Therefore $\mu_4 = m(n^2-1) \left\{ \frac{(m-1)(n^2-1)}{48} + \frac{(n^2-4)}{80} + \frac{1}{48} \right\};$

$$\beta_2 = \frac{3(m-1)}{m} + \frac{9(n^2-4)}{5m(n^2-1)} + \frac{3}{m(n^2-1)}$$

$$= 3 - \frac{6}{5m} - \frac{12}{5m(n^2-1)},$$

an expression which tends to the value 3 as m and (to a lesser extent) n increase. Since x is a discontinuous variable, a correction of $\frac{1}{2}$ is applied; and the normal deviate is therefore

$$\text{N.D.} = \frac{|\text{Total Rank Value} - \frac{1}{2}m(n+1)| - \frac{1}{2}}{\sqrt{\frac{m(n^2-1)}{12}}}.$$

The effect of tied rankings has not been considered. However, if ties are present, the true variance will be slightly less than that given above for the appropriate m and n . Thus the normal deviate obtained by this procedure will be slightly smaller than the true value, and the corresponding probability slightly too great. If a significant value is obtained when tested against the uncorrected variance, the presence of ties in the rankings will merely mean that the result has slightly more significance than the calculated figure would suggest.

IV. EXAMPLE

Eight judges rank seven objects. The experimental object has a total rank value of 43. What is the probability that this or a higher value would be obtained by chance?

Variance $= \frac{8}{12} (49 - 1) = 32;$

Normal Deviate $= \frac{43 - \frac{1}{2} \times 8(7+1) - \frac{1}{2}}{\sqrt{32}} = 1.8556;$

$P = .0318$ (approx.).

To obtain the exact value from the tables we have to reverse the direction of the deviation, as the other tail only is tabulated. The corresponding deviation in the tables is therefore $m(n+1) - \text{Total Rank Value}$. In the example this is $64 - 43 = 21$. From the table for $n = 7, m = 8$ we find that a total rank value of 21 or less would be obtained by chance in .03113 of the trials. The approximation agrees closely with this value.

REFERENCE

Kendall, M. G. (1948). *Rank Correlation Methods*. London: Griffin and Co.

Distribution of Total Rank Value for One Particular Object in m Rankings of n Objects

TABLES OF EXACT VALUES OF P FOR TOTAL RANK VALUES

The probability values given below are for deviations in the predicted direction only, and should be doubled if absolute probability values are required.

$m = 2$

Total Rank Value	n					
	3	4	5	6	7	8
2	.11111	.06250	.04000	.02778	.02041	.01563
3	.33333	.18750	.12000	.08333	.06122	.04688
4		.37500	.24000	.16667	.12245	.09375
5			.40000	.27778	.20408	.15625
6				.41667	.30612	.23438
7					.42857	.32813
8						.43750

$m = 3$

3	.03704	.01563	.00800	.00463	.00292	.00195
4	.14815	.06250	.03200	.01852	.01166	.00781
5	.37037	.15625	.08000	.04630	.02915	.01953
6		.31250	.16000	.09259	.05831	.03906
7		.50000	.28000	.16204	.10204	.06836
8			.42400	.25926	.16327	.10938
9				.37500	.24490	.16406
10				.50000	.34111	.23438
11					.44606	.31641
12						.40625
13						.50000

$m = 4$

4	.01235	.00391	.00160	.00077	.00042	.00024
5	.06173	.01953	.00800	.00386	.00208	.00122
6	.18518	.05859	.02400	.01157	.00625	.00366
7	.38272	.13672	.05400	.02701	.01458	.00854
8		.25781	.11200	.05401	.02915	.01709
9		.41406	.19520	.09722	.05248	.03076
10			.30400	.15895	.08746	.05127
11			.43200	.23920	.13578	.08057
12				.33565	.19783	.11987
13				.44367	.27280	.16968
14					.35860	.22974
15					.45190	.29907
16						.37598
17						.45801

TABLES OF EXACT VALUES OF P FOR TOTAL RANK VALUES—*continued*

The probability values given below are for deviations in the predicted direction only, and should be doubled if absolute probability values are required.

 $m = 5$

Total Rank Value	n					
	3	4	5	6	7	8
5	.00412	.00098	.00032	.00013	.00006	.00003
6	.02469	.00586	.00192	.00077	.00036	.00018
7	.08462	.02051	.00672	.00270	.00125	.00064
8	.20988	.05469	.01792	.00720	.00333	.00171
9	.39506	.11730	.04032	.01620	.00750	.00385
10		.21680	.06916	.03241	.01499	.00769
11		.34863	.13824	.05877	.02749	.01410
12		.50000	.21984	.09799	.04683	.02417
13			.32224	.15201	.07479	.03912
14			.43904	.22145	.11287	.06018
15				.30517	.16202	.08844
16				.39969	.22241	.12476
17				.50000	.29321	.16962
18					.37234	.22304
19					.45683	.28436
20						.35229
21						.42493
22						.50000

 $m = 6$

					.058	.054
6	.00137	.00024	.00006	.00002	.00006	.00003
7	.00960	.00171	.00045	.00015	.00024	.00011
8	.03841	.00684	.00179	.00060	.00071	.00032
9	.10700	.02050	.00538	.00180	.00178	.00080
10	.23045	.04980	.01344	.00450	.00393	.00176
11	.40329	.10254	.02918	.00990	.00785	.00352
12		.18433	.05649	.01968	.01453	.00655
13		.29590	.09907	.03588	.02517	.01143
14		.42920	.15994	.06076	.04111	.01893
15			.23968	.09647	.06378	.02991
16			.33606	.14463	.09448	.04529
17			.44397	.20585	.13423	.06601
18				.27939	.18349	.09293
19				.36310	.24207	.12671
20				.45358	.30898	.16772
21					.38252	.21595
22					.46035	.27093
23						.33176
24						.39713
25						.46538
26						



Distribution of Total Rank Value for One Particular Object in m Rankings of n Objects

TABLES OF EXACT VALUES OF P FOR TOTAL RANK VALUES—*continued*

The probability values given below are for deviations in the predicted direction only, and should be doubled if absolute probability values are required.

$m = 7$

Total Rank Value	n					
	3	4	5	6	7	8
7	.00046	.00006	.00001	.0 ⁵ 4	.0 ⁵ 1	.0 ⁶ 5
8	.00366	.00049	.00010	.00003	.00001	.0 ⁵ 4
9	.01646	.00220	.00046	.00013	.00004	.00002
10	.05167	.00732	.00154	.00043	.00015	.00006
11	.12986	.01971	.00422	.00118	.00040	.00016
12	.24691	.04492	.01005	.00283	.00096	.00038
13	.41015	.08936	.02125	.00610	.00208	.00082
14		.15820	.04070	.01206	.00416	.00164
15		.25305	.07162	.02209	.00775	.00307
16		.37012	.11686	.03787	.01359	.00543
17		.50000	.17824	.06122	.02260	.00915
18			.25574	.09388	.03584	.01477
19			.34714	.13716	.05445	.02293
20			.44794	.19170	.07954	.03432
21				.25717	.11205	.04972
22				.33216	.15259	.06987
23				.41421	.20137	.09543
24				.50000	.25802	.12693
25					.32161	.16466
26					.39065	.20864
27					.46315	.25856
28						.31377
29						.37329
30						.43586
31						.50000

$m = 8$

8	.00015	.00002	.0 ⁵ 3	.0 ⁶ 6	.0 ⁶ 2	.0 ⁷ 6
9	.00137	.00014	.00002	.0 ⁵ 5	.0 ⁵ 2	.0 ⁶ 5
10	.00686	.00069	.00012	.00003	.0 ⁵ 8	.0 ⁵ 3
11	.02393	.00252	.00042	.00010	.00003	.00001
12	.06447	.00743	.00127	.00029	.00009	.00003
13	.14129	.01854	.00327	.00077	.00022	.00008
14	.26078	.04033	.00750	.00178	.00052	.00018
15	.41564	.07805	.01555	.00379	.00111	.00038
16		.13638	.02959	.00745	.00222	.00077
17		.21768	.05210	.01369	.00415	.00144
18		.32034	.08573	.02369	.00736	.00259
19		.43826	.13263	.03887	.01242	.00443
20			.19392	.06071	.02007	.00727
21			.26918	.09065	.03113	.01152
22			.35622	.12983	.04654	.01763
23			.45115	.17888	.06724	.02616
24				.23771	.09406	.03770
25				.30540	.12769	.05289
26				.38017	.16852	.07233
27				.45953	.21655	.09656
28					.27135	.12602
29					.33203	.16095
30					.39728	.20139
31					.46543	.24714
32						.29772
33						.35237
34						.41012
35						.46982

NOTES AND CORRESPONDENCE

ON MAKING STATISTICAL ASSUMPTIONS

By A. S. C. EHRENBERG

Institute of Psychiatry, Maudsley Hospital

In psychological circles there are several schools of thought which regard mathematical and statistical methods, particularly if at all complex, with distrust. The criticisms appear usually to run along logical as opposed to empirical lines. Zangwill (1), for example, bases his criticisms on Chambers' pronouncement that 'it is at least very doubtful whether the concept of measurable quantity may be applied at all to psychological qualities' (2); and we are repeatedly told that the question of whether or not psychological qualities are measurable is a matter of logical or formal 'assumption.' At the same time the commonest line of attack on numerical methods in psychology is the complaint that their use involves too many 'assumptions.' Clearly, therefore, the nature of such so-called 'assumptions' and the possibility of error either in overlooking or in relying on them requires some examination, even without explicitly entering into the wider issues involved.

At least three kinds of error can occur in setting out the assumptions on which an argument is based. First, some of the assumptions implicit in the argument may be omitted or insufficiently explained. Secondly, the nature of the relevant assumptions and of their logical status may be misinterpreted. Thirdly, so-called assumptions may be introduced which cannot strictly be regarded as assumptions at all.

To illustrate the need for a closer scrutiny, I shall refer more particularly to the recent volume by Professor Harold Gulliksen on *Theory of Mental Tests* (3), and the comments passed on it in a review published in the *British Journal of Psychology* (General Section, XLIII, pp. 80-1) over the initials E. G. C.¹ Here the reviewer marshalled a set of 'assumptions' requisite for the application of statistical procedures in a manner which is admirably concise and so lends itself readily to a brief examination. In the review printed in this *Journal* (V, pp. 134-5) it was stated that in Gulliksen's book "the assumptions and definitions are carefully formulated": and this opinion also, as will be seen from what is said below, seems much too generous.

1. Failure to give *all* the assumptions inherent in an argument is a well-known fault; and failure to stress and explain the most important assumptions is still commoner. This type of error can be most simply illustrated by a familiar problem, namely, the use of normal distributions. Thus in Chapter XIV of Professor Gulliksen's book we find that the requirement of normality for Wilks' tests is not mentioned at all. More serious is the fact that Gulliksen emphasizes that almost all his formulæ are true, irrespective of any assumptions about the distributions involved. But the touchstone in applied mathematics is not so much whether the formulæ are 'true' (that the psychologist may leave to the mathematician), but whether they can in fact be legitimately applied. Now, unless we assume that the variables discussed by Gulliksen—'true' scores, 'error' scores, etc.—are in fact normally distributed, it is impossible to make precise and valid numerical deductions about the behaviour of our data from the formulæ set out by him. Normality is an inherent assumption in almost all correlational work (particularly that arising out of the theory of mental tests); but it is an assumption which must be fulfilled, not so much to make the formulæ true as to make them useful. Admittedly Gulliksen briefly mentions ((3), p. 17) a need for assuming normality distributed *errors*, but he makes no reference to the distribution of the *raw* data. This rough and ready way of dealing with the distributional problem appears particularly unfortunate, since in practice many observations are not normally distributed. Indeed, the more advanced investigator often seems to do his level best to make the *joint* distribution of his data *not* normal by adopting systematic experimental designs (e.g., in working with an 'experimental' and a 'control' group).

At times there seems to be a peculiar difficulty in detecting the more isolated omissions. When, for instance, in the two reviews mentioned above, it is stated that the assumptions in Gulliksen's book "are scrupulously pointed out whenever they are made" or that they have been "carefully formulated," the reviewers may have been misled by the fact that *certain* assumptions have been stressed into supposing that *all* assumptions have been mentioned.

2. When assumptions are explicitly stated, there is often some confusion in the way their nature or their status is treated. Not infrequently it seems to be implied that they have to be accepted by an act of faith. In the review already quoted, for example, certain statistical assumptions are specified, and the writer then asks: 'How many psychologists would be *willing* to make them?'

¹ The editor of the *General Section*, to whom my comments were originally submitted, suggested that "it might be better to state the arguments more fully in the *Statistical Section*."

Once again examples can be found in the discussions of normality. Thus, in regard to the so-called 'error' scores, the reviewer tells us that 'without this or some similar assumption further progress cannot be made.'¹ But this normality is not an 'assumption' that the practical worker can make it merely by being 'willing' to do so. What is needed is an empirical confirmation: e.g., the data should be first fitted by some mathematical distribution function, possibly normal, possibly not. Such a procedure, however, turns the issue into a matter of descriptive statistics, not of simple assumption. If the data cannot be adequately described by normal distributions, then theoretical formulae which assume normality should not be used.

Again, it is asserted that tests are sometimes 'assumed' to be parallel, i.e., to have equal means and equal standard deviations. In such cases it should be the preliminary duty of the investigator to determine whether the tests are 'parallel' (in this sense) or not. As before, the notion of being 'willing' just to assume such parallelism really indicates a serious omission in the analysis.

It seems possible that this mistaken attitude about 'having to make assumptions' may be a symptom of quite a different difficulty that is never explicitly stated. The reviewer's anxiety about assuming parallelism, for instance, might be due to his concern over the fact that in psychometric theory there seems to be no sound procedure for estimating the reliability of non-parallel data. If this is so, it indicates that a criticism, which at first sight might appear to be mere semantic quibbling, is needed to direct attention to points of obvious practical importance.

3. Perhaps the most dangerous examples of so-called 'assumption' are to be found in those statements which either cannot be verified or else are true by definition. When a writer claims to base an argument on certain assumptions, he really implies that his conclusions are capable of proof; whereas in the cases to which I am referring either they cannot be proved to be true (thus apparently requiring a 'willingness' to make the requisite 'assumption') or else they follow directly from the definitions and consequently are not in need of proof. It is perhaps instructive that much of the inconclusive work in psychological statistics—notably in arguments where a mathematical model is set up—may be traced to a confusion of this kind. As an instance, may I refer once again to the review in the *General Section*, where the writer refers to "the basic assumption that a person's raw score . . . is composed of a 'true' score and an 'error' score"? As it stands, it would be impossible to show that this statement is either true or untrue. We might perhaps amend it to read: "Let us try to describe the data by a model involving two variables, which may be called 'true' and 'error' scores, and may be defined as follows . . ." But the specification of a model is in itself not a matter of assumption or proof. For much the same reasons it is misleading to say that the 'true' and 'error' scores are assumed to be uncorrelated: such a statement would in effect be part of the definition of 'true' and 'error' scores in the model proposed.

The change in formulation is clearly of practical consequence. Instead of asking whether the mathematical model proposed is 'true' (in some way not explicitly defined), and assuming its 'truth' for the sake of the subsequent deductions, it becomes the investigator's business first to ascertain whether the model set up is internally consistent, and then whether it furnishes an appropriate scheme for describing the data.

Finally, three general points may be briefly considered. First, how is the concept 'assumptions' to be defined? Secondly, what is the effect of the foregoing arguments on the reviewer's evaluation of Professor Gulliksen's book? And lastly, how far is it legitimate to generalize the arguments here illustrated?

To attempt a formal definition of what is meant by the term 'assumption' is to invite criticism from various quarters. Yet some such definition seems needed, at least for *statistical* assumptions. We may begin by excluding broad and general axioms or postulates about the nature of empirical reality or about the bases of our logical systems: disagreement about these, though interesting and suggestive, rarely affects what the practical scientist does. Secondly, we may exclude scientific hypotheses as such, that is to say, undemonstrated propositions which are only provisionally put forward with a view to testing their plausibility in the light of the data. I suggest, therefore, that an 'assumption' may be defined as a proposition which may or may not be true in fact, but which must hold good if a given theoretical result is to be validly deduced. An obvious instance arises in the derivation of some formula incidentally required for estimating a parameter or testing significance. Some of the commonest statistical assumptions of this type assert the normality, linearity, or homoscedasticity of the distributions to which the formulae are to be applied. In empirical science all such assumptions must in principle be verifiable; for, if one cannot discover whether some such incidental assumption is plausible in any given case, one cannot judge the plausibility of the final inference.

The verdict on Professor Gulliksen's book, pronounced by the reviewer in the *General Section*, declares that 'psychological research workers will . . . turn aside with feelings akin to horror.' Since it is stated that 'the algebra is worked out clearly enough,' the reviewer's reasons must presumably arise from his doubts over the assumptions made. Now, although on this latter count his argument

¹ He appears to be in error when he describes Gulliksen as assuming that the *standard deviation* of the 'error' scores is normally distributed (cf. (3), p. 17).

seems unacceptable, nevertheless his rejection of the book might still be endorsed for quite the opposite reason, namely, that the algebra and the statistical concepts appear, on closer scrutiny, to be decidedly confused.

Lastly, if it be agreed that the contentions here advanced are sound so far as they go, it is tempting to suggest that the main points may reasonably be generalized. Certainly, it seems likely enough that the widespread distrust of statistical methods in psychology is really due, not to the profounder logical or formal difficulties that are commonly said to beset all attempts at numerical measurement, particularly as these difficulties are seldom explicitly analysed; rather it would appear that it chiefly springs, however indirectly, from difficulties of a purely technical or empirical nature. Many psychometric techniques, for instance, prove to be extremely restricted in their applications, and in practice these restrictions are constantly forgotten. Others turn out to be inherently faulty as judged from a technical standpoint.

REFERENCES

1. Zangwill, O. L. (1950). *An Introduction to Modern Psychology*. London: Methuen & Co.
2. Chambers, E. G. (1943). 'Statistics in Psychology and the Limitations of the Test Method,' *Brit. J. Psychol. Gen. Sect.*, XXXIII, 191.
3. Gulliksen, H. (1950). *Theory of Mental Tests*. New York: Wiley.

ON MAKING STATISTICAL ASSUMPTIONS: A REPLY

Mr. Ehrenberg's *Note* on making statistical assumptions, which might perhaps more aptly be described as a review of a review, calls for a few comments. This is perhaps scarcely the place to enter on a controversy about assumptions. But, as the pilloried reviewer, I may say that considerations of space allowed me only 200 words in the *General Section* of the *British Journal of Psychology*, in which to describe Professor Gulliksen's book and express an opinion on its value to psychologists. Mr. Ehrenberg uses more than six times as much space in the present issue of the *Statistical Journal*, and apparently comes to the same conclusion—that the book is not to be recommended. Anyone who, distrusting my review, reads Gulliksen's book for himself would very soon discover the ambiguities in exposition (such as the one indicated in Mr. Ehrenberg's second footnote).

I believe that psychologists understood me when I used the word 'assumption' in the dictionary sense of 'something taken for granted.' Mr. Ehrenberg's claim that "one must always be able either to verify or disprove" assumptions does not find a place in the definitions of the *Oxford English Dictionary*; and it seems that he has forgotten that statistical methods are based on the principles of probability ('the calculus of uncertain inference') and that nothing can be proved or disproved by statistics.

E. G. C.

THE SCIENTIFIC STUDY OF PERSONALITY

A Note on the Review

By H. J. EYSENCK

Institute of Psychiatry, Maudsley Hospital

Reviewers have a great deal of latitude in the way in which they express their opinions; and no author should be too sensitive when he finds his most dearly cherished thoughts given short shrift. But criticism must be of what the author has actually written; it must avoid misquotations, factual errors, and other objectively fallacious arguments. The review of my book *The Scientific Study of Personality* (see this *Journal*, V, pp. 208f.) deviates so far from what is permissible that a note of protest appears necessary. I shall take up the reviewer's errors in numbered series, as they form an almost perfect example of the art of distortion.

1. *Change of sense through direct misquotation.* On p. 211 he prints the following sentence: "Burt's system of classification," he (i.e., Eysenck) says, "is some thirty years out of date, and discredited in contemporary psychology for many valid reasons." What I actually say (p. 41) is: "Burt's work is based on the rating of the strength of McDougallian 'instincts' in children by teachers. The choice of a system of classification some thirty years out of date . . ." No one reading the passage in its context can doubt that it is McDougall's system which I consider out of date; the quotation is subtly changed in such a way as to give a completely erroneous impression.

2. *Change of meaning through indirect misquotation.* In the same paragraph, the reviewer, referring to Burt's 1915 *Report*, says that "Burt tells us, the correlations were based on 'items of behaviour systematically recorded.'" This is incorrect. Burt does not tell us at all what his correlations were based on. He merely discusses different possible ways in which ratings could be obtained, and which of these methods he considers superior.

3. *Change of meaning through incomplete quotation.* The reviewer quotes my sentence "Burt . . . has contrasted the factor of 'emotionality' with the factor of 'neuroticism'" (p. 115). He goes on to say that this is incorrect, and that Burt in fact considers the two to be "virtually identical." However, I continue my paragraph by quoting Burt's very words, bringing out clearly in what way he contrasts the two factors. Had the reviewer not broken off the quotation where he did the reader could have judged for himself the accuracy of my interpretation.

4. *Disparagement through imputation.* He discusses my claim that objective tests are superior to other types of tests. Because I do not quote the recent Symposium in the *Brit. J. Educ. Psychol.*, he concludes that I am "apparently unaware of the fairly comprehensive survey of the reliability and validity of different methods there brought together." He does not appear to consider the possibility that other causes than ignorance may be responsible; and that other psychologists might not agree that the survey was "fairly comprehensive." It was not quoted because it was in my view so one-sided in its presentation, and so restricted in its coverage, as to be practically valueless for my purpose. To construe this into ignorance on my part of its very existence is unworthy of a scientific review.

5. *Disparagement through suppression and indirection.* The reviewer criticizes my use of objective tests by saying that the usual objection to such tests is that their reliability is low, and that their validity cannot possibly exceed their reliability. He complains that my tables seldom include reliability coefficients, and that those I give are by no means high. The tests specifically mentioned by him—dexterity, persistence, tapping, body-sway—are reported in the literature to have reliabilities which are quite satisfactory; body-sway suggestibility test-retest reliability, for instance, is above .9, tapping is above .8, dexterity is between .7 and .9, while persistence has an estimated minimum reliability of .7. These values are not particularly low; they would almost certainly be higher than reliabilities of the methods advocated by Burt and the reviewer. Surely the main point is that I have laid stress throughout my book on validity, rather than on reliability, which is merely a tool for the purpose of reaching higher validity. He is right in saying that an unreliable test cannot be valid. Surely the conclusion is that as these tests have been shown to be valid, they must *eo ipso* be reliable? He suppresses the well-known fact that these tests have high reliabilities, and uses indirection to suggest that because they are believed (by him) to be unreliable, therefore they cannot be valid—suppressing the fact that they have been shown to be valid, and must therefore be reliable.

6. *Misrepresentation through falsification of argument.* On p. 210, he deals with my treatment of the problem of whether normal and pathological phenomena are continuous or not. He says: "Dr. Eysenck . . . prefers to assume that" they are continuous, and "having decided to assume continuity" goes on from there. The reader of the review would never guess from this that far from assuming continuity the main purpose of the experiments described in the book was to decide this question of continuity, and that I made no assumption at all about what I would find. I believe that the main contribution I have made on the experimental side to the study of psychiatry has been precisely this attempt to find conclusive evidence on this point as between the qualitative point of view and the quantitative one. It is possible to argue that I failed in this attempt; my critic prefers to ignore it altogether, and to say calmly that I "assume" continuity. This is a complete misrepresentation of the situation.

7. *Misrepresentation through invention.* On p. 209 he says that in order to secure a unique and

invariant position I carry out "a rough graphical rotation." In fact the rotation is precise and algebraic. I do not quite see how one could carry out a graphical rotation under the special circumstances even if one wanted to do this. The "rough graphical rotation," therefore, is entirely an invention of the reviewer's: it finds no counterpart in my book.

8. *Misrepresentation through ignorance.* In dealing with criterion analysis, he says: (having quoted my aim in rotating as that of reaching a position for the factor-axis where the correlation between factor and criterion is a maximum): "one would have supposed that in doing this the investigator would eventually approach a position which was identical with the criterion-axis, and that the maximized correlation . . . would be simply the ordinary multiple correlation between the criterion and the tests." This quotation shows that the reviewer has failed to understand the theory of criterion analysis. There can be no multiple correlation between the criterion and the tests, because the criterion is not included in the factor analysis, and cannot be so included, by virtue of the fact that the analysis is carried out *separately* on the normal and the abnormal groups, *not* on the *combined* groups. This is the most important point in criterion analysis. If the reviewer has failed to appreciate the consequences that follow, it may be legitimate to query whether his understanding is adequate to the task he has undertaken.

9. *Misrepresentation through carelessness.* He follows the criticism quoted above by saying: "It is true that he (Eysenck) goes on to say that one may also increase the correlation by improving the criterion. . . . How this is to be done we are never told." The reader is told at length on pages 76 and 77. W. L. G. would be justified in arguing that he did not agree with the arguments—or did not understand them. He cannot plead that they are not to be found.

10. *Misdirection through non sequitur.* The reviewer quotes very briefly my account of the hypothetico-deductive method, as involving the postulation of an hypothesis, the deduction of certain consequences, and the verification of these consequences along experimental lines. He then declares it to be invalid by quoting Judge Jeffreys' saying: "If Mistress Gaunt had been guilty, you may rest assured she would have been burned. She was burned. *Ergo*, she was guilty." In some obscure way, he seems to feel that this story is relevant, and disproves my own argument. This seems to me a complete *non sequitur*. I leave it to the reader to decide whether he regards this as a telling and relevant argument.

11. *Misrepresentation through misstatements of facts.* In connexion with my discussion of scientific methodology he says: "To prove his (Eysenck's) hypotheses, it is not sufficient to report experiments that are consistent with his corollaries: *he should also have attempted to disprove the alternatives.* When he does notice alternatives, they are usually dealt with by the short and easy method of 'abusing the defendant's attorney.'" When we look at the facts, we find that the main theory tested, namely, that of the qualitative or quantitative nature of abnormal mental states, as compared with normals, is examined by means of a crucial experiment, i.e., an attempt is made to deduce consequences from either theory which would be incompatible with consequences deducible from the other. The critic may argue that the attempt was not successful; he can hardly remain consistent with the facts of the case and maintain that I never even made the attempt to disprove the alternative hypotheses. Similarly, I do not try to invalidate current psychiatric diagnostic systems by abuse. I deduce certain consequences (such as inter-psychiatrist reliability) which must follow from a satisfactory system, show that these consequences in actual fact are not to be found, and conclude that the original hypothesis is not tenable in its present form. The arguments may be faulty; but it is contrary to fact to say that there is 'nothing but abuse.'

12. *Use of alternative plea.* The alternative plea is very useful in law, where a man may plead that he did not kill the victim, but that, if he did, he did it in self-defence. Scientific criticism tends to frown on the use of this device, which is exemplified in its purest form by my reviewer. On p. 210 he assures the reader that my treatment of abnormal mental states as quantitatively different from normal people is opposed to Burt's assumptions and methods; on p. 211 he assures us that the two systems and the concepts involved are similar, and that I have in effect taken mine from Burt. But, while either one or the other of these arguments may be true, they cannot both be true. There are other instances of this peculiar type of argument in the review; it would hardly be worth while to exume them.

13. *Misrepresentation through wrong emphasis.* Throughout his review, the writer attempts to relate my results to those of Burt, castigates me for not agreeing with Burt, attempts to establish Burt's priority with respect to such concepts as *emotionality*, and quite generally writes as if no other psychologist but one of the editors of the *British Journal of Statistical Psychology* had ever written on the subject of personality. This emphasis is erroneous in two ways. In the first place, my book is an account of experiments, not an historical treatise: occasionally previous writers are mentioned, but there is no systematic attempt to cover the ground. This was originally contemplated; but the material was too extensive, and will soon appear as a separate publication under the title *The Structure of Human Personality* in the Methuen series. Criticisms of an historical character are appropriate to this later volume: they are irrelevant to the present one. In the second place, the criticisms made are poorly taken. If there are similarities between Burt's system and mine, which I would be the last to deny, they presumably are due largely to the fact that the early workers (Jordan and Gross on the theoretical side, Heymans and Wiersma on the experimental side) laid down the

foundation on which we both have built. I had already commented on the relation between my system and Burt's in *Dimensions of Personality*; there seemed little point in discussing it again. The reviewer seems to regard this lack of repetition as proof that I have changed my mind—a curious hypothesis also put forward by Sir C. Burt. In fact, the very term "emotionality," with almost exactly the connotation which it carries in Burt's system, was used by Heymans and Wiersma in 1909; and the existence of the trait so denoted verified by an empirical procedure closely analogous to factor analysis. If, therefore, I should have entered into the historical side, the correct emphasis would not have been on Burt, but on these other, earlier workers who anticipated his system (and mine) in considerable detail. But it is not with respect to the system itself that I would claim to have made any great advance: my aim has been rather that of improving the underlying methodology and making more precise the measurement of the dimensions posited.

Here, then, are a baker's dozen of serious misrepresentations of fact, of misquotations, of errors, of misunderstandings, and of complete inventions having no counterpart in the book under review. The list could have been extended, but it is perhaps long enough to cast doubt on the reviewer's knowledge of the field, his ability to understand the complex arguments involved, and his impartial fairness in assessing the validity of proofs and demonstrations.

THE SCIENTIFIC STUDY OF PERSONALITY: A REPLY

Methodology. In answering Dr. Eysenck's thirteen criticisms I shall assume that readers will be more interested in the appraisal of his own methods and conclusions than in personal questions about my competence as a psychiatrist or psychologist, which, I modestly admit, is sadly defective. If I have in any respect misrepresented or misquoted his views, I offer at once my sincerest apologies. Nevertheless, a reference to what I actually wrote will, I think, show that he has misquoted me far more often than I have misinterpreted him.

He has told us that the chief aim of his book was to "improve the underlying methodology" in the field of personality research. It was with this that my review was chiefly concerned; and it is with this that my reply will chiefly deal. What he describes as 'incomplete quotation,' 'misrepresentation through suppression,' 'non-sequiturs,' and the like, arose mainly from my efforts to be brief. Hence the best way to meet his complaints will be to amplify the points over which misunderstanding seems to have arisen.

The Verification of the Proposed Hypothesis. I began, as he says, by "quoting very briefly" the three steps in his 'account of the hypothetico-deductive method'—"postulation," 'deduction,' 'verification.' So far, I gather, there was no 'distortion.' But "then," according to him, I "declare it to be invalid" by just narrating a fallacious syllogism attributed to Judge Jeffreys. "In some obscure way," says Dr. Eysenck, "the reviewer seems to feel that this story disproves my (Eysenck's) argument." If, however, Dr. Eysenck will read the paragraphs and the footnote that follow "this story," he will see that it was inserted merely to illustrate the formal nature of the argument, not to disprove it. I went on to indicate quite explicitly why such a mode of inference was defective, namely, that it "commits the fallacy of affirming the consequent." In deductive logic the only way to prove a hypothetical premiss is to disprove its contradictory (Stebbing, *loc. cit.*, p. 105). Instead of elaborating the point at greater length, I thought it sufficient to refer to several logicians and statisticians who had already discussed it, and whose objections Dr. Eysenck had ignored in his book, and still ignores in his reply.

In his chapter on the subject, instead of dealing with the criticisms so often raised, he preferred to appeal to a well-known authority, and for this purpose introduced the long quotation from Miss Stebbing's text-book (*A Modern Introduction to Logic*, 1930). But this mode of defence suffers from two grave weaknesses. First, Miss Stebbing's chapters can hardly be regarded as the last word on statistical methodology. Secondly, her own discussions do not support his view; and there can be little question that Dr. Eysenck has seriously misrepresented, or at least misunderstood, her real opinion.

If the reader will refer to the original, he will be amazed to find that Dr. Eysenck, who now lays such emphasis on exact quotation, has run together in one continuous extract sentences and passages scattered over many pages of her book, rearranged their order, reworded several of the statements, and even inserted a sentence of his own.¹ "The method of science" (he makes her say, though she says no such thing) "... has four essential stages"; and then he gives her analysis, not of 'the method of science,' but of 'common sense thinking.' She herself points out that an induction proceeding by these four stages is invalid, since "a formal fallacy is involved in the inference that a given hypothesis is true because the consequences deduced from it have been verified"; and explains in some detail how the "method of science" improves on this inconclusive procedure in several "important respects."

¹ If the reader will compare pages 227 to 242 of Miss Stebbing's book with pp. 16-18 of Dr. Eysenck's he will find excellent instances of 'misquotation,' 'incomplete quotation,' 'misrepresentation through suppression,' 'wrong emphasis,' 'invention,' etc., etc., all in a single passage. I will not go so far as Dr. Eysenck, and suggest that it offers "an almost perfect example of the art of distortion"; but it certainly presents a very misleading picture of Professor Stebbing's own analysis of inductive methods as applied to the physical and the biological sciences.

The Disproof of Alternative Hypotheses. The obvious way to escape any such formal fallacy would be to attempt a direct disproof of all the conceivable alternatives to whatever hypothesis the writer wishes to maintain. In empirical science, however, the most we can seek will be, not deductive proofs leading to conclusions that are *certain*, but arguments based on probability, leading at best to an inference that the hypothesis favoured by the investigator is more *probable* than the remaining alternatives.¹ I thought it unnecessary to amplify these points in any detail because they had already been adequately dealt with in the references I quoted.

But although in section 10 of his reply Dr. Eysenck seems to reject my general contention, in section 11 he apparently admits it. He now claims that, "for the main theory tested," he did in point of fact "disprove the alternative hypotheses." But he interprets this requirement in a way very different from most other methodologists. Hence I still think my two chief criticisms were justified. First, the need for this all-important step is nowhere mentioned in his account of the 'hypothetico-deductive methodology.' Secondly, in what are described as the central enquiries of the book, there is no *direct* attempt to disprove the alternatives. We are, it is true, told that, where Kretschmer holds a different view, his methodology is 'faulty' and his attempts at proof 'disingenuous.'² For the rest the disproof of the chief alternative is treated as an incidental consequence of the proof of the main hypothesis, whereas the correct mode of reasoning would invert the order. However, in view of the way in which he seems to have missed my principal point, let us examine a little more closely his study of psychosis, since this is the part of his book which he complains that I ignored, and the part which has chiefly aroused criticism from medical writers. Here he claims, by means of the hypothetico-deductive procedure, to have established two far-reaching conclusions.

Hypothesis I. The first is the hypothesis that "mental states diagnosed as psychotic by psychiatrists" are merely instances of extreme variation in a single *quantitative* continuum—the 'dimension of psychoticism,' which ranges from perfect normality to definite psychosis. The alternative would be the view that certain psychoses at least differ *qualitatively* from each other and from the normal state, much, for instance, as Korsakow's psychosis is ordinarily supposed to differ from general paralysis, and both from normality. In accordance with the steps described in his opening chapter, Dr. Eysenck first deduces what he takes to be the consequences of the 'quantitative hypothesis' (that of continuity), and then attempts to verify these. But, so far as I can discover, he deduces no consequences from the *alternative* hypothesis, nor does he undertake any independent investigation to disprove them.

To begin with, he puts all his cases of psychosis³ into a single group. No attempt is made to show that his '100 psychotic subjects' were typical psychotic patients; indeed, from their excellent performances at the tests most psychiatrists would be tempted to suppose that they were highly selected cases. He then argues that, if the psychotic group is continuous with the normal, tests which differentiate significantly between the two should yield *positive inter-correlations for each group separately* (his italics). He adds that these inter-correlations should be "proportional to the power of each test to differentiate between the normal and psychotic groups," the 'differentiating power' being measured by a biserial correlation.⁴ He offers no formal demonstration that this corollary follows logically from his hypothesis: he merely "points it out." The coefficients obtained from his tests are tabulated in Tables XL and XLI; and he thereupon concludes (p. 219) that "we find our deduction verified, and it would appear that psychotic states do in fact

¹ It is in regard to the relation between induction and probability that Miss Stebbing's section on 'The Problem of Scientific Method' seems somewhat out of date. The reader may usefully compare both her views and Dr. Eysenck's with those put forward in Russell, B., *Human Knowledge*, 1948, esp. Pt. VI, 'The Postulates of Scientific Inference.'

² In his *Note* Dr. Eysenck ends his section on this point by declaring: "it is contrary to fact to say that there is 'nothing but abuse.'" But I said no such thing. What I did say was that, in discussing alternative theories, he seemed a little too apt to "abuse the plaintiff's attorney." I used this familiar cliché to mean that, instead of adducing facts to disprove the views he dislikes, he relied rather too freely on an attempt to disparage the scientific competence of those who advocate some alternative conclusion.

³ In attempting to assess the real implications of Dr. Eysenck's statistical studies, the practical psychiatrist is rather at a loss owing to the inadequate description of the patients tested. We are simply told that his "psychotic subjects" (as he calls them) were "divided into two equal groups consisting of 50 manic-depressives and 50 schizophrenics." No details are vouchsafed to show whether the individuals so labelled were typical of the highly heterogeneous conditions to be found under these somewhat vague diagnoses. I should have thought that by now even the statistical psychologist was convinced of the need for combining case-studies with mere quantitative analysis. His use of biserial correlation to measure the 'discriminative power' of his tests presupposes that the frequency distributions for the combined group (e.g., the 100 normals and the 100 psychotics) would be, not only continuous, but also normal (as shown in his schematic diagrams). Now, if we reduce the two constituent groups to appropriate proportions, we can test the assumption of normality by applying Pearson's equations for the effect of selection on the correlations and standard deviations. So far as can be judged from the data Dr. Eysenck gives, it seems clear that the hypothesis of a normal distribution along the single continuum is far from being fulfilled.

form a continuum with normal." A few paragraphs later he adds: "we have shown" that psychotics are not "qualitatively different from normal people," i.e., that the alternative hypothesis is disproved. But the only evidence for this further inference is a parenthetical assertion (unproved) that the consequences of his own hypothesis "would not follow on any tenable counter-hypothesis." Apparently he assumes that with any 'counter-hypothesis,' e.g., the hypothesis that differences were qualitative, the average inter-correlation would be zero. Yet this is never proved; and is it really so? Suppose we were investigating the analogous alternatives in a study of sex-differences: would Dr. Eysenck expect to find that if sex differences were qualitative (due, say, to an additional X-chromosome), then the inter-correlations between measurements for height, weight, shoulder breadth, etc., would yield an average correlation of zero in males and females taken separately?

Or consider the parallel question that arises in the field of cognitive psychology: do patients in mental hospitals who have been diagnosed as imbeciles or idiots differ from normal persons solely because of some "quantitative variation in a single continuum"? Take any two tests that will 'differentiate significantly' between 100 normals and 100 defectives, say, any pair from Abelson's battery; and calculate their correlations within each group taken separately. The coefficients will undoubtedly be, not zero, but positive and significant. On taking all the tests in the battery, we should find that the average of the correlations for each (and therefore its factor-saturation) would vary in close proportion to its 'power to discriminate' between the normals and the defectives. Dr. Eysenck would thereupon conclude that the results had shown that "neither [idiots] nor [imbeciles] are qualitatively different from normal people." Nevertheless, among institutional cases of imbecility, as Professor Penrose assures us, while a small proportion display merely an extreme quantitative variation in general intelligence, as many as 75 to 80 per cent. suffer from some "chronic neuropathological condition," due to some specific gene, or to some disease, deformity, or qualitative defect (often quite specific) without which they would have had "average intelligence."

The fact is, as an algebraic proof would quickly show, that Dr. Eysenck's assumed 'deductions' do not follow from his hypotheses as he states them. In any case, his decision to rely almost exclusively on graded tests scarcely affords a genuine chance of proving or disproving the presence of a qualitative difference. In explaining his choice of test procedures, he asserts that individuals "cannot be meaningfully said to differ unless they are compared along some quantitative variable." But, as I went on to argue, this is tantamount to assuming the very continuity that he later claims to demonstrate. In reply he now protests (section 6) that he did not make any assumption about what he would find. Naturally I did not mean that he deliberately or consciously did so. Yet at times he certainly speaks as if the main object of his investigation was merely to confirm, by a quantitative factorial technique, an assumption which, on more general grounds, he took to be almost a foregone conclusion from the start. Thus, in his preliminary remarks on 'the dimensional approach,' he tells us that the whole notion that persons can be sorted into qualitative categories—a view "often maintained by some psychiatrists"—really rests on an obsolete type of thinking. "Such thinking," he says, "has always characterized early stages of scientific progress: usually it is shown later that what was considered qualitatively different is reducible to quantitative variation along a continuum."¹ No doubt, as he contends, unless some such postulate is accepted, factor analysis as he understands it could hardly be applied; and (so he maintains) "at the present stage of our knowledge there does not exist any other method which could aid us in our quest." But that is a somewhat sweeping and surprising pronouncement; moreover, it is a pronouncement which (as I said in my review) most psychiatrists would assuredly deny. There are many other methods available, observational, experimental, etiological, and clinical, as well as factorial and statistical. And even among factorial and statistical procedures there are several which would appear to be just as appropriate for qualitative characteristics as Dr. Eysenck's are for quantitative variables.

Hypothesis II. I now turn to the second of the two main hypotheses about the nature of psychoses which Dr. Eysenck has discussed. As he points out, the views of Kretschmer and other psychiatrists imply the presence of at least one further 'dimension' or factor, namely, "a continuum running from the extreme schizophrone to the extreme cyclothyme" (or manic-depressive). Dr. Eysenck's twenty tests are designed to investigate this second hypothesis as well as the first. Their inter-correlations do, as a matter of fact, yield two factors. Of these the first is identified with the factor of psychoticism; and therefore, it is argued, the second should be identical with the dimension contrasting 'schizophrenics' with 'depressives,' if such a dimension exists.

¹ Here it is surely Dr. Eysenck who is basing his own generalization on views that are "thirty years out of date." What he says is not true of other sciences. "The idea of a continuum seems so simple to us," says Schrödinger, "owing perhaps to conditioning in early childhood. . . . If, however, we envisage the development of physics during the last half century, we get the impression that the discontinuous aspect of nature has been forced upon the scientist *much against his will* (Schrödinger's *Time*, 1951, pp. 30, 54f.) The same is true in the biological sciences, notably in the genetic study of human abnormalities. "The view that ascribed all differences to continuous variation has got to be abandoned: what is fundamental are qualitative differences, not quantitative" (cf. Myerson, A., *The Inheritance of Mental Diseases*, 1923).

Now the saturations for the second factor reveal a zero correlation with the biserial correlations which measure the power of the tests to discriminate between 'schizophrenics' and 'depressives.' Therefore Kretschmer's notion of a second psychotic factor is disproved: "it would appear that schizothymia-cyclothymia does not exist as a separate dimension of personality." Moreover, since the scores of the schizophrenics are in general somewhat better than those of the manic-depressives, we may (it is suggested) reasonably conjecture that schizophrenia is, as it were, a mild form of manic-depressive insanity, intermediate between the latter and perfect normality.

But surely there are other ways of accounting for the zero correlation and the scores. First of all, just because Eysenck's second factor and Kretschmer's second factor are both orthogonal to psychoticism, it does not follow that the two must be one and the same; several dimensions or factors might exist, all orthogonal to psychoticism and to each other. Secondly, the saturations for Eysenck's second factor are of very doubtful value. They are reached after a double reversal of signs; and, whenever the initial coefficients are small, such reversals are extremely arbitrary: with a different mode of reversal, quite a different correlation with the biserial coefficients might have been obtained. In any case, the figures correlated have for the most part a decidedly low degree of significance; of the biserial coefficients, at least 16 are definitely non-significant; of the factor-saturations none of the negative figures in the normal group is significant, and two at most in the psychotic group. Thirdly, it is quite possible that such tests as 'reading prose,' 'abstractions,' 'mirror drawing,' and the like—even if their reliability remained reasonably high when performed by psychotic patients—were not suited to elicit the special characteristics of the schizothymic and cyclothymic temperaments. In view of the similarities between schizothymic tendencies and introversion, and between cyclothymic tendencies and extraversion, I should have thought that the methods adopted with some success by previous investigators in their endeavours to verify the factor of introversion and extraversion might have furnished more convincing evidence. The mere fact that "the requisite deduction could not be verified" hardly justifies the far-reaching inference that is drawn. Finally, the test results themselves seem highly inconclusive. Dr. Eysenck's first and most important factor ('psychoticism') only accounts for about 10 per cent. of the test-variance. His second factor is smaller still. Indeed, he himself does not venture to say what it really represents.

I do not doubt that most of the mental characteristics which Dr. Eysenck has tested, and many of those which, in an exaggerated degree, form the more superficial symptoms of the neuroses and milder psychoses, might be expressed in terms of quantitative factors of the usual type. But that is by no means the same as "showing that neither neurotics nor psychotics are qualitatively different from normal people." The temperature, rate of pulse, rate of breathing, and other measurable symptoms of patients suffering from measles or typhoid fever could be expressed in terms of one or two quantitative factors. But that does not demonstrate that there are no qualitative differences between measles and typhoid, or between either of these and normal health.

Reliability of Tests. In my review, I ventured to suggest that possibly the low coefficients which Dr. Eysenck has obtained—biserials, inter-correlations, and saturations—might be attributed to the fact that the types of test chosen were not likely to yield very reliable measurements, particularly with neurotic and psychotic patients, and I thought it unfortunate that Dr. Eysenck's tables "seldom give reliabilities." In his note he replies that this is "disparagement by suppression." Nevertheless, in his thirty tables I can still discover only two which mention 'reliabilities,' and these relate to researches carried out by other investigators. He now tells us that four of the tests he himself has used (body-sway, persistence, and the like) "are reported in the literature to have satisfactory reliabilities"; but he gives no references. The references that I myself suggested (which give rather low reliabilities for tests of the general type in question) he dismisses as "onesided." In any case, he says, the reliabilities for his own tests (or rather their lower limits) could be inferred from the validity coefficients. But with the twenty tests used to discriminate psychotics from normals the highest validity coefficient was 0.402, and with those used to discriminate neurotics from normals none reached 0.600: the averages are about half the size. Such values can scarcely be claimed as proving that the reliabilities are "satisfactory."

Dr. Eysenck roundly declares that I have "suppressed the fact that these tests have been shown to be valid, and must therefore be reliable." I did not 'suppress' the fact; I questioned whether

¹ In accordance with the principles of 'criterio-analysis,' Dr. Eysenck also examines the correlations obtained for certain factors reached by rotation. Once again he finds them non-significant; and therefore considers his conclusion to be confirmed. But in any case mere rotation could not possibly discover any new dimension.

In section 7 of his reply, he complains that I wrongly supposed that his method of rotation was graphical, whereas it was "precise and algebraic." If that was a mistake, I apologize. However, in the initial exposition of the proposed procedure (p. 73), he says nothing about an 'algebraic' rotation, and he does speak of "a graphic demonstration of the results." He dismisses my remarks on 'criterion-' or 'criterio-analysis' (as he variously terms it) by saying that my 'understanding (of it) is inadequate.' But my main criticisms were really those urged by his former collaborator Dr. Lubin (*Psychol. Rev.*, LVII, pp. 54f.), who must surely have understood the procedure. To these Dr. Eysenck still attempts no reply.

it was a fact. And at this point a critical reader might be tempted to ask whether it is not Dr. Eysenck who has "suppressed" some of the relevant facts. In his book (p. 125) he states that the "testing was carried out by Drs. Clarke and Gravely" (who at that time were co-operating with him at the hospital), and that "the account given by Clarke has been followed . . . often verbatim." But Dr. Clarke tells us that "none of the tests" (which included "body sway" and others "selected in consultation with Dr. Eysenck as the most promising") "has a sufficiently high validity coefficient to claim any practical value for screening or diagnosis" (*Brit. J. Educ. Psychol.*, XX, p. 202). Dr. Gravely reports very similar results: (see the theses quoted by Dr. Eysenck). These findings scarcely support his present line of defence.

Priority. Dr. Eysenck's last and longest section is devoted to the question of priority. He complains that "throughout [my] review [I] attempt to relate [his] results to those of Burt, castigate him for not agreeing with Burt, attempt to establish Burt's priority with respect to such concepts as 'emotionality'" (a concept which, he says, was really due to Heymans and Wiersma), and generally that I "write as if no other psychologist had ever written on the subject of personality." The last sentence is surely an instance of what Dr. Eysenck has labelled "misrepresentation through misstatement of fact." If he will be good enough to refer once again to my review, he will see that I refer to more than a dozen other psychologists by name, including "eminent writers on personality, like Allport . . ." and several others whom Dr. Eysenck himself had ignored. I made no attempt to establish anyone's priority in regard to the concept of emotionality. The sole priority with which I was concerned was priority relative to Dr. Eysenck, and that in respect of his general scheme of factors, not in respect of one particular concept. Nor did I "castigate" him for not agreeing with Burt; I simply regretted that, on the numerous points about which he *did* agree, not with Burt merely but with other investigators, the only reference he generally gave for the "demonstration" of the several factors was, as a rule, some particular paper of his own. He replies that his work was not meant to be "an historical treatise." But I was not asking for a history, but simply for a less misleading mode of reference. His "bibliography" runs to more than 300 numbers; yet this, like the text, omits most of the previous British investigations which have dealt with the factors he describes and reached much the same conclusions. I had presumed that the omissions might be due to the fact that he was unfamiliar with the earlier English literature. He rejoins that apparently I did not "consider the possibility that other causes than ignorance might be responsible." I did; but thought it more charitable to dismiss them.

He professes surprise that I should relate his scheme to that of Burt. There are two obvious reasons. First, apart from the single factor of psychoticism and a slight change in nomenclature, the two schemes seem almost identical. Secondly, Dr. Eysenck received his own psychological training under Burt, and carried out his earlier factorial researches in his laboratory. Indeed, in his previous book he had proposed "a theory of personality organization," which, he said, was "based on Burt's views of the nature of the statistical factors and on the hierarchical theories of McDougall."

However, he now explains that, "if there are similarities between Burt's system and his," they must be due chiefly to the fact that both he and Burt built on the foundations laid down by early workers like Jordan, Gross, Heymans and Wiersma. As early as 1909, he tells us, Heymans and Wiersma used "an empirical procedure" which (in his view) is "closely analogous to factor analysis" in order to demonstrate the existence of hypothetical traits like emotionality. This explanation of the resemblances, I confess, is somewhat unexpected. Nowhere in his book is any of these four names mentioned. Nor do I know anything to support the view that Burt derived either his methods or his hypotheses from the work of these four writers rather than from that of Wundt, McDougall, and Pearson. In point of fact, as a glance at the dates will show, his first investigation with factorial methods was started under McDougall some two or three years before the date given by Dr. Eysenck for the publication by Heymans and Wiersma. Dr. Eysenck's first investigation on the subject (based, as he expressly states, on Burt's factorial technique) was not published until 30 years later.

All psychiatrists would readily admit that, at any rate until quite recently, German contributions to the study of mental disorders had been well in advance of British. But I venture to think that Dr. Eysenck's enthusiasm goes a little too far when he claims that the origins of the factorial study of personality are also to be found in the *Zeitschrift für Psychologie*. Nevertheless, it is well to be reminded of the work of early continental writers, which British writers have perhaps too often overlooked.

Misquotation. The rest of the thirteen charges in Dr. Eysenck's long indictment will perhaps seem to the reader too trivial to spend time over. However, since they have been pressed so strongly, it is only right that my answers should be briefly recorded. The more important relate to alleged misquotations. But in most cases the misquotation appears to be Dr. Eysenck's rather than mine. Thus, in section 2 he refers to 'Burt's 1915 Report' describing the general and specific emotional factors discovered by analysing correlation tables; and in commenting on his account I had stated that the coefficients were "based, not on impressionistic ratings obtained from teachers, but on items of behaviour systematically recorded." But here, he says, I am guilty of misquotation, since Burt "does not tell us on what his correlations were based." If so, I apologize to Professor

Burt; but the '1915 Report'¹ explicitly states that the investigation "departed from the method hitherto used in studies of character qualities" and that the assessments correlated were based on items of behaviour "systematically observed and recorded for all members of the groups." It is therefore Dr. Eysenck whose account conveys 'a completely erroneous impression.'

In section 3 he gives yet another example where my method of quotation is said to distort what was meant. On page 115 of his book Dr. Eysenck had asserted that, in a chapter contributed to the American Symposium, "Burt has . . . contrasted the factor of 'emotionality' with the factor of 'neuroticism.'" I ventured to correct this, and observed that Burt's real contention was that the two are "virtually identical." Dr. Eysenck now tells us that my quotation was inaccurate and broke off too soon. But, as a matter of fact, it is Dr. Eysenck who breaks off the quotation too soon. Burt's essential point is that the difference lies, not in the factors themselves, but only in their names. He cites an enquiry carried out among both neurotics and normals which showed that the so-called general factor of 'neuroticism' was not confined to neurotics only, but was also found among those who were perfectly healthy; and from this he concludes that the factor is "little else than the factor of general emotionality."² Dr. Eysenck, in quoting this paragraph, omits this concluding sentence.

Although throughout his *Note* Dr. Eysenck directs his criticisms mainly to Burt's supposed opinions, it should be observed that many other factorists³ have come to much the same conclusions as those expressed in my review. It is thus Dr. Eysenck who writes as though only one psychologist of repute had put forward opinions differing from his own. Indeed, he goes out of his way to of ascribe every antagonistic view to the same single source, often (to borrow his phrase) by "subtly changing" what I actually wrote. In section 12 of his *Note*, for example, he contrasts my discussion of his 'hypotheses' with my discussion of his 'conclusions,' and alleges that at the beginning of the former I "assured the reader that [Eysenck's] treatment of abnormal mental states as quantitatively different from normal people is opposed to Burt's assumptions." But I said nothing at all about Burt's assumptions in regard to adult psychotics, which, for all I know, may be much the same as Dr. Eysenck's. What I said was, "many psychiatrists" would be opposed to Eysenck's assumptions, and would probably reply that such differences "were to be expressed in terms, not of a single quantitative variable, but of qualitative attributes." I supposed that anyone familiar with the literature of this controversy would know that I had in mind such views as those put forward most convincingly perhaps by Eliot Slater in various papers (e.g., 'The Neurotic Constitution,' *J. Neurol. Neurosurg. Psychiat.*, VI, 1943, pp. 1-16, 'A Heuristic Theory of Neurosis,' *ibid.*, VII, 1944, pp. 49-55) and tacitly or explicitly adopted by all but a few psychiatric text-books.

Let me add that, although I have ventured to criticize the accuracy of his arguments, this does not mean that I myself necessarily agree with the views that he has rejected. Moreover, I hope that his numerous complaints about the way he supposes I have 'disparaged' his work will not blind readers to the fact that my review went on to express a sincere admiration of "its undoubted merits"; and I more than once paid a tribute to the "ingenuity and industry it displays."

What I most of all regret in his reply is that, by concentrating on these rather trifling points about the incompleteness of the arguments or quotations in a review that was already too long, he has diverted attention from the more important issues I endeavoured to stress. I had hoped that the questions I raised might serve to clarify the conflict between Dr. Eysenck and the psychiatrists (to which Professor Lewis refers in his Foreword) and perhaps elicit a fuller and clearer explanation

¹ I take it that this refers to his 1915 L.C.C. Report. But the method used is also indicated in the brief abstract of his British Association paper. If Dr. Eysenck still has doubts on the matter, he will find the experiment fully described in Burt's later article in *Character and Personality*, and the method itself is again explained in *Brit. J. Med. Psychol.*, XVII, p. 164. Let me add that, in trying to meet Dr. Eysenck's criticisms and verify his quotations, I have found myself constantly handicapped by the careless way he gives his references. In his bibliography, for example, the only contribution given for Burt under the year 1915 is a paper read to the British Association; but, if one turns up the volume and page specified by Dr. Eysenck, one finds that there is nothing whatever by Burt on that page at all. Passing to the next reference to discover which of Burt's 1937 papers is being criticized in section 1, the reader finds that the only article cited under that year is one which was not published in that year. And there are many similar slips which cannot all be due to the printer. ² *Feelings and Emotions* (ed. M. L. Reymert), 1950, p. 548. Elsewhere Burt had contrasted the so-called general factor of neuroticism with the various specialized factors for certain types of neurotic disorder. The former he believes to be wrongly named, and to have diverted attention from the latter, which, he suggests, are from a psychiatric standpoint far more distinctive (see this *Journal*, I, pp. 193f.).

³ See the numerous references cited by Raymond Cattell in his *Description and Measurement of Personality* (Index, s.v., 'General Emotionality Factor' and 'Neuroticism'). Cattell himself writes: "Probably the worst example in applied psychology of confusion through inadequate factoring occurs with respect to the alleged 'neuroticism' factor; later studies show half-a-dozen such factors, and the first has no more claim than the others to be 'the neuroticism factor.'" He himself identifies it with his 'factor C.' This, he says, 'ties up with general emotionality,' but is as important in the causation of delinquency as in the causation of neurotic disturbances: (cf. also *id.*, *Personality*, p. 502).

of the statements that had aroused some criticism. Unfortunately in his *Note* as in his book Dr. Eysenck merely waves away the views of those who differ from him as being 'one-sided,' 'valueless,' or 'based on lack of understanding or knowledge of the field'; and we still have no real reply to the questions themselves.

In criticizing his 'attempt to improve the underlying methodology,' my real objection was not so much that his factorial technique might be faulty in this or that detail, but rather that a purely factorial approach is of itself very liable to oversimplify the facts. What I was pleading for was a more eclectic 'methodology,' like that, for example, which has proved so successful in "unravelling delinquency"—a set of co-operative researches which should combine the method of case-study with that of statistical analysis, and the method of clinical observation with that of the laboratory test. And since Dr. Eysenck seems to imagine that I am just an isolated psychiatrist, unfamiliar with the requirements of a strict scientific or statistical procedure, and that my "knowledge of the field" is confined to just one brand of factorial psychology, let me add that, since I wrote, others have brought forward much the same criticisms.¹ With a far better knowledge both of the opportunities and of the difficulties, they have also urged what I ventured to hint at towards the close of my review: namely, that, placed as he is on the staff of one of the best-equipped mental hospitals in the country, it seems disappointing that he has not been able to make fuller use of the unique facilities thus afforded for linking up his tests and quantitative data with the first-hand clinical observations made by the skilled psychiatrists who form his colleagues and who presumably had charge of the psychotic patients he has tested. What might be achieved along these lines has already been foreshadowed in the papers contributed by the little band of psychologists and psychiatrists who worked together before the war at the same mental hospital under the guidance of Professor Spearman. And I should like to echo the remarks of Dr. Stephenson, one of the most active members of that group: "by themselves factorial studies can serve only the broadest aims; conjoined with case-records, each can help to interpret the other, and together reveal common characteristics and common trends."

W. L. G.

LABORATORY NOTES

To the Editor, *British Journal of Statistical Psychology*

SIR,—In previous publications in your *Journal* and elsewhere I have seen references to 'Laboratory Notes' on (1) the Psychology of Humour and (2) the Content Method of Scoring the Rorschach Tests. If copies are still available, we should be grateful to know how they can be procured. If not, would it be possible to have some indication of the contents?—Yours, etc., E. A. TELFORD.

Reply.—A few copies of the Notes are still available, and can be obtained by writing to me at the Department of Psychology, University College, Gower Street, London, W.C.1. The note on 'Humour' includes a fairly complete summary of the theories of humour, and a report of illustrative investigations carried out in the laboratory (down to 1936, when the note was prepared) dealing with the more conspicuous points of disagreement. Tables of correlations are given both 'between items' and 'between persons,' together with the results of the factorial analyses (based mainly on work by Pelling, Dewar, and Woods). The note on the Rorschach Test tentatively puts forward a detailed scheme for deriving trait assessments from a study of the contents of the replies rather than of their form (see this *Journal*, III, pp. 24f.).—CHARLOTTE BANKS.

PUBLICATION OF TABLES

SIR,—I venture to suggest that authors who wish to contribute statistical papers to any of the journals published by the British Psychological Society should be required to supply copies of the essential tables on which their inferences are based. If possible, the tables should be printed with the article. If (for reasons of expense, space, or lack of general interest) this is impracticable, then the copies might be deposited for reference in the Society's office.

To illustrate the difficulties arising when tables are not available, I may refer to a recent paper on retardation among school pupils, published in another journal. The author gives no tables of correlations or saturations; but describes in some detail the factors obtained apparently by some method of rotation. We happen to be carrying out a research in the same field at the Queen's University, Belfast; and in this case I was very courteously allowed to consult the unpublished thesis. On examining the tables, it seemed clear that another interpretation might easily be placed on the loadings obtained. Other readers might desire to compare these factors with those published by Barakat, Sutherland, and other authors in this *Journal*. And in many cases it might be difficult, if not impossible, to obtain access to such theses. I therefore wish to put forward this proposal as a general policy for all editors who publish articles based on statistical data.—Yours, etc., JACK WRIGLEY.

¹ Dr. O'Connor perhaps has put the matter most emphatically. "Dr. Eysenck's false god," he writes, "is the statistical table... so pleasing to that kind of scientific mind which prefers to remain aloof from the human element" (*Bull. Brit. Psychol. Soc.*, III, 1952, p. 115). Dr. Maddox has criticized the tests employed even more severely than I had done; and several later reviews, I find, raise much the same objections as I had.

BOOK REVIEWS

Contributions to Mathematical Statistics. By R. A. Fisher. New York: John Wiley and Sons, Inc. London: Chapman and Hall. 1950. Pp. 680. £3.

This large volume contains the first comprehensive collection of Sir Ronald Fisher's writings on mathematical statistics, and forms a notable addition to Wiley's series of statistical publications. The selection has been made by the author himself. Forty-two contributions have been taken from articles and lectures published before 1943; a further note refers to three others, not actually included, of which one appeared as early as 1915. The papers, some of which were printed in octavo journals, others in quarto, have been reproduced on their original scale by a photographic process. Professor Fisher has made a number of deletions and emendations, and provided brief introductory notes, linking the various contributions with each other and setting them in the context "of the common knowledge, or common misapprehensions, current at the time they were written." An index to the whole has been prepared by Professor S. S. Wilks.

The modern theory of mathematical statistics has been almost entirely created during the last seventy years. The *Contributions* thus cover nearly half this period. Although in many respects they form a fairly self-contained body of knowledge, it is scarcely possible to understand the achievement they represent, except in relation to statistical developments during the earlier half of the same period. This witnessed the appearance of the long series of publications by Professor Karl Pearson and his 'biometric school'; and it might be said that the place held in the earlier phase by Pearson's 'Mathematical Contributions to the Theory of Evolution' and kindred papers is as central as that held by Fisher's *Contributions* in the later. Under Pearson's leadership, detailed investigations were made of frequency functions; these led to the well-known Pearsonian classification of frequency curves and the use of moments for fitting them to empirical distributions; most of the multivariate descriptive procedures appropriate to biometry, including nearly all the familiar coefficients of correlation and regression, were systematically developed; and the problems of sampling, and of significance-tests as applied to large samples, were studied both experimentally and mathematically. The chi-squared distribution—the first of the more modern frequency distributions—was rediscovered. The chi-squared distribution has since become closely associated with Pearson's name. On the other hand, the nature of statistical inference was very imperfectly understood. The method of inverse probability, with Bayes' postulate, was accepted as the foundation; routine procedures were regularly based on the calculation of probable or standard errors. The famous phrase 'more than three times the probable error' (or 'twice the standard error'), which has figured so conspicuously in the literature of experimental psychology, was almost the hallmark of statistical analysis at that time. A classic instance, of special interest to the psychologist who of late has proved the chief user, is that of the tetrachoric correlation. This was but one of many coefficients devised by Karl Pearson for describing degree of association in a 2×2 table. For its probable error he proposed a complicated formula as long ago as 1913. Yet to the present day no exact sampling distribution is assumed for a correlation so computed. Thus, not even if an underlying normal bivariate distribution is assumed do we possess any rigorous test of significance for this coefficient. Indeed, prior to 1915 the exact sampling distribution of the product-moment coefficient itself had not been determined, though both Gosset ('Student') and Soper had given approximations.

Nevertheless, while the general problem of statistical inference remained confused, the line of progress had been indicated as early as 1908 when Gosset, in his classical paper 'On the Probable Error of the Mean' (*Biometrika*, VI, pp. 1 *et seq.*), published the *t*-distribution. On the other hand, the theory of statistical estimation remained extremely backward. The importance of 'degrees of freedom' was entirely unrecognized; and thus the essential foundations of modern statistical analysis had still to be laid.

It was shortly after this that Professor Fisher made his first important contribution by solving the outstanding problem of the exact distribution of the correlation coefficient (*Biometrika*, X, 1915, pp. 507-21). This was followed a few years later by two articles in *Metron* 'on the probable error of a coefficient of correlation deduced from a small sample' (which incidentally gave the exact sampling distribution of the intra-class coefficient) and on 'the distribution of the partial correlation coefficient.' The substance of these three articles forms the subject of the note which now makes up Paper 1 of the present volume: since "the established policy of *Metron* does not permit republication of articles from that journal," the papers themselves are not reproduced. The 1915 paper contained the first use of Professor Fisher's celebrated 'geometrical method' (differing from earlier methods by the novel device of representing a sample of n variates as a single point)—a method which he brought to bear with such elegance and effect on various problems of sampling. In the second of

these articles we have the origins of the z -distribution and of the analysis of variance. In a biographical note (published in *Sankhyā*, 1938, and included in the introduction to the volume) Professor Mahalanobis discusses how Fisher's early training and practice in the use of geometrical methods may have been partly responsible for his subsequent applications of them; but he offers no clue to why Professor Fisher should have chosen to work with z , i.e., with half the natural logarithm of the variance ratio, rather than with the variance ratio itself, which was subsequently tabulated by Snedecor and labelled, by way of a pleasant compliment, with the initial F .

The papers which follow are presented for the most part in chronological order of publication. But here and there this is abandoned for a grouping according to subjects. The reasons for the mixed classification are not made clear. One would have thought that a subject classification should either have been used consistently throughout or else not introduced at all. As it is, some of the articles referred to in earlier sections do not appear till a later place in the volume, and no dates are given in the table of contents. Thus the subject of Paper 2 (1920), which contains the concept of a sufficient estimate, leads directly on to that of Papers 10 (1922) and 11 (1925), in which the problem of estimation is systematically discussed. The late insertion of Paper 10 is particularly confusing in view of the references to maximum likelihood that occur in earlier papers. In between we find Paper 3, which deals with a problem in the analysis of time-series and describes the use of orthogonal polynomials as "a means of isolating groups of components susceptible of separate explanation." This topic is then resumed in Papers 16 and 37 which deal with harmonic analysis. Incidentally we may observe psychologists, in spite of the frequent occurrence of time-series in researches on curves of work or learning, seem to have made little use of these methods.

Papers 4 to 9 have chi-squared as a common theme, all except the first being concerned with "goodness of fit." Before these articles were published it was the practice to take one less than the number of frequency groups as the quantity defining the appropriate chi-squared distribution for the test of significance, no matter what the form of the data. But here we are introduced to the now familiar concept of *degrees of freedom*. Paper 6 (1922) gives the distribution of regression coefficients, and is particularly rich in its implications. The t test is shown to provide the exact test for all regression coefficients 'linear or non-linear, simple or multiple,' and the z distribution, which, "before its general applicability was recognized, . . . kept turning up unexpectedly," is shown to be the appropriate distribution for testing the fit of regression lines, instead of the older chi-squared distribution. It is, however, unfortunate that the prefatory note to Paper 6 should read "Its [the z distribution's] relationships were first summarized (Paper 8) in 1924," when Paper 12, in the same year, seems evidently intended.

Paper 12 provides an admirable demonstration of the relations between the distributions of t , z , χ^2 , and a normal variate. The following paper (13) should also be of value to students, since it was written "to give mathematical teachers a concise account of the mathematical distributions used in the common tests of significance"; and the next paper (14) describes the general sampling distribution of the multiple correlation coefficient. Since very few statistical text-books written for psychologists offer a clear account of the relations between these different distributions, Professor Fisher's systematic discussions are especially to be recommended.

Papers 17 and 18 are concerned with problems which formed the starting point for Professor Fisher's book on *The Design of Experiments*; several later papers (39, 40, and 41) have to do with various mathematical ramifications of this subject. It is only during recent years that these principles have begun to influence the statistical and experimental procedures adopted (not always appropriately) by research workers in psychology.

Another series of papers, numbered 22, 25, 26, and 27, include the more important of his contributions to the problems of statistical inference. No one who employs tests of statistical significance can safely omit a proper study of the principles involved. The special features of Professor Fisher's treatment are, first, an unqualified rejection of Bayes' postulate, and, secondly, a strong emphasis on randomization rather than sampling as a condition for the valid use of statistical tests. However, even statisticians have found his discussion of these problems hard to follow; and the position is made no easier for the psychological research-worker by the presence of a rival approach (based on sampling and the theory of 'two types of error') which has been developed by Neyman and Pearson. The issues are primarily logical and may therefore in a large measure be appreciated without any advanced mathematical knowledge. Paper 27 (on 'Uncertain Inference') provides a comparatively easy approach: it examines the issues from an historical standpoint, contains no mathematics, and leads to a clear statement of Professor Fisher's own views. Of the remaining papers those of greatest interest to the psychologist are numbers 32, 33, and 34, which deal with the discriminant function—a device which has already been used in psychological investigations and has formed the subject of articles published in this *Journal*.

In view of its size and scope the volume can scarcely be considered expensive. It is primarily a work for the specialist. Yet it contains much that is of the highest value for all who aspire to an understanding of statistical methods rather than of mathematical statistics. Arthur Summerfield

Progress in Clinical Psychology. Vol. I, Section 2. Edited by D. Brower and L. E. Abt. New York : Grune & Stratton. 1952. Pp. xxiv + 266. \$5.00.

In their preface the editors announce that they have planned to issue every second year a volume summarizing the progress of clinical psychology. The first in the series, however, is intended to provide "as complete a coverage as possible of the past six years." Section 1, already published, presented "an historical and systematic résumé of developments" and a detailed account of "diagnostic and therapeutic instruments and procedures." Section 2 is divided into four parts, dealing respectively with developmental processes, applications to special areas, approaches to clinical psychology, and professional issues. Each chapter is written by a specialist in the branch, and includes a detailed appendix of references.

In studying the various summaries, the point that is most likely to impress readers of this *Journal* is the way in which the current developments in America are tending to stress the value of statistical research and particularly of quantitative checks on both theoretical and practical claims. As Professor Cattell observes, "American clinicians seem now to realize that their best hope of making substantial progress in clinical problems lies in the development of refined statistical procedures." The longest chapter in the volume is Professor Kogan's review of 'Recent Statistical Methods' as developed for clinical work. He deals with four main topics.

(i) First, he tells us, "one of the most active areas during recent years has been the attempt to devise systematic methods for handling profiles or score patterns in ways acceptable to the clinician." After reviewing the more important devices that have been proposed, he concludes that those likely to prove of the greatest service are the techniques of correlations between persons, of intra-class correlation, and of multiple discrimination. One interesting application of the principle of 'correlating persons,' the 'so-called Q-sort technique,' is being tried at the Counselling Center at the University of Chicago, suggested, it appears, by Dr. Stephenson: on the other hand, we are told, "Dr. Cattell has criticized both the nomenclature of the method and its dependence on non-objective data." Dr. Cattell himself contributes a separate chapter on P-technique. Intra-class correlation has the advantage of well-established tests of significance, but seems to have been used less freely in America than over here. In illustration of the possibilities of the discriminant functions Dr. Kogan refers to work reported in earlier articles in this *Journal*, particularly by Lubin, Rao, Slater, and others, in discriminating neurotic and normal groups.

(ii) A related field of work arises out of the need to develop special methods for validating tests of personality, since those hitherto used with tests of cognitive abilities have been dismissed by many clinical workers as inappropriate. Several investigators, notably Zubin, have put forward devices for validating projective techniques, while Cronbach has proposed a rather complex procedure (based on Vernon's blind matching procedure) for validating qualitative assessments of personality.

(iii) A third line of study that is attracting considerable attention is the attempt to develop criteria for assessing 'purity of measurement,' e.g., Guttman's method of scalogram analysis, Lazarsfeld's theory of latent structure, and the various factorial methods for studying the dominance of the general factor. Professor Kogan declares that he himself is "prejudiced towards prognosticating that the scaling approach (not necessarily linear scaling) will be found to meet the requirements more adequately than the approach by multiplex testing, but the scaled components must be synthesized back into reasonable models of human beings."

(iv) The 1952 symposium on 'Research Design in Clinical Psychology' (as well as other papers cited by Professor Kogan) has attracted increasing attention to the need for a more scientific planning of investigations in this field. Professor Kogan deplores the fact that so far he has been "unable to discover any comprehensive discussion of the logic of research designs comparable to that of Greenwood in the field of sociological research." The value of a control group has been increasingly recognized, especially in assessing the results of different modes of treatment. But he considers it disappointing that hitherto so little use has been made of Fisher's proposal to "substitute the multi-factorial experiment for the single factor method to secure a broader basis for inductive generalizations." In Great Britain, largely owing to the direct influence of Professor Fisher himself, the 'latest refinements in the analysis of variance' have been more promptly and more widely used. "Recent developments," so Dr. Kogan believes, "portend a closer integration of analysis already done in this factor analysis." Here, however, he appears to have overlooked the work already done in this direction by Bartlett, Kendall, and those British factorists who have attempted to "integrate" the work of Fisher and of Pearson: as it happens, the very first issue of this *Journal* (1947, I, pp. 3-26) contained a systematic attempt to show the mathematical relations between analysis of variance on the one hand and analysis of covariance on the other. Dr. Kogan rightly points out that clinical investigators are far too ready to declare that their most pressing problems are unamenable to the stock statistical techniques, without considering whether some of the newer devices may not furnish the very method which their problem demands. Among more recent developments the merits of

so-called "non-parametric or distribution-free methods of statistical inference," such as the 'order statistics' of Wilks and the ranking methods of Kendall, rightly receive considerable stress.

In his chapter on 'P-technique Factorization,' Professor Cattell begins by urging a greater use of statistical methods generally. For clinical research he, too, considers the most useful to be factor analysis and the analysis of variance, since they do not require that artificial control of the variables, or that artificial elimination of all variables except one, which has proved so difficult to achieve in dealing with human beings. Factor analysis he regards as "the ideal device for dealing with the total personality"; and in particular, he maintains, "the new factorial procedure known as P-technique¹ gives precisely an analysis and description of unique personal traits."

In the remaining chapters the surveys of the literature on special branches of clinical psychology may enable the reader to judge the uses made of statistical methods in this or that field by the summaries they give of the latest results. Their general effect, it may be said, is to demonstrate the need for greater caution in regard to the somewhat optimistic claims so often put forward in the field of clinical psychology during the last few decades.

The publication of these biennial surveys will be welcomed in this country as warmly as in the United States. It is therefore unfortunate that no very systematic attempt seems to have been made to cover recent British contributions, especially as several of the methods and results quoted as new by the American authors have long been accepted over here. There are generous references to work published in this *Journal*, but practically none to relevant articles in the other psychological journals published in England. Nevertheless, in their selection of material and in their appraisal of results, the general editors and the authors of the several chapters display a discernment that is at once sound, critical, and impartial. Certainly, if we may judge by the two sections now issued, the series will be of the greatest value to all students of clinical psychology, and should be a challenge to the statistical psychologist to produce and apply more appropriate research-techniques in this important field of work.

Cyril Burt

The Young Delinquent in his Social Setting. By T. Ferguson. Oxford University Press, 1952. Pp. xii + 158. 10s. 6d.

In an earlier volume (*The Young Wage-Earner*, 1951) Professor Ferguson has described an enquiry into the after-histories of 1,349 Glasgow boys who left the ordinary schools at 14. The present study forms a supplement to that broader survey. Its object is to determine the frequency of juvenile delinquency and to ascertain its apparent causes. With a view to obtaining further light on physical and mental factors, minor studies were also made of smaller groups from schools for the physically or mentally handicapped.

He finds that over 12 per cent. of the boys from the ordinary schools had been convicted of crime before the age of 18. Among these "a low level of scholastic ability was probably the most powerful single factor associated with a high incidence of crime." Unfortunately it was not possible to test their intelligence; and the reader may wonder whether the real factor was, not so much the low level of scholastic attainments, but rather the low level of innate capacity, which, as earlier enquiries have shown, is a common characteristic of juvenile delinquents and the main cause of educational backwardness. No attempt was made to assess emotional or temperamental qualities or to determine whether these may not be even more important than intellectual.

In common with most psychological investigators, Professor Ferguson comes to the conclusion that physical factors by themselves have little influence, but that environmental conditions play a considerable part. Of these the most important, he believes, are frequent change of employment and residence in an overcrowded district. The influence of 'broken homes,' he finds, is not so great as other writers have supposed: "in good housing districts the proportion convicted was below the average, whether the family was 'broken' or not; in bad housing districts the proportion was above average, and slightly worse in families which were not 'broken.'" Among those who were physically handicapped "the absence of a parent" proved less important than "the presence of a bad parent."

If a youth lives in a home with three or more persons to a room the risk of his becoming delinquent is about five times as great as if he lived in a room with less than three. Professor Ferguson infers that "the abatement of crowding of a severity of three persons or more to a room would apparently produce a reduction in the amount of juvenile delinquency equal to about one-fifth—provided that in

¹The term is used to denote 'the factorization of an individual case' (i.e., what I myself had referred to as 'intra-individual variability': cf. Burt, C., 'Correlation between Persons,' *Brit. J. Psych.*, XXVIII, 1937, p. 60). Professor Cattell explains that there is a tendency to extend Stephenson's term 'Q-technique' to cover correlations between different persons, and therefore "it would be convenient to adopt Burt's term P-technique to refer to this analysis, within a single person," i.e., what he himself had elsewhere called 'intra-individual covariation' (cf. Cattell, R., *Description and Measurement of Personality*, p. 99).

their new setting the families conform to standards of conduct prevailing among families not at present severely crowded." But will they conform? That surely is the crux. By the title he has chosen Professor Ferguson tacitly invites a comparison between his methods and results and those set forth in the volume on *The Young Delinquent* by Professor Burt. Now Burt has shown that, to a very large extent, delinquency is a by-product of a general inferiority, characterizing families at the lower end of the socio-psychological scale and apparently resting on an innate and inheritable basis. Intellectual, educational, temperamental, moral, and civic inferiority are all correlated together; and such families tend to create or drift into the unhealthy overcrowded districts where delinquents are commonly found. Improvement in living conditions might do something towards solving the problem, particularly in the case of those individuals in whom extrinsic conditions are more important than intrinsic; but it could hardly cut down delinquency to the extent he implies.

Professor Ferguson's conclusions might be summed up by saying that, so far as the conditions covered by his survey are concerned, the factors conducing to youthful delinquency appear after all to be much the same in Glasgow as in London and other large cities where surveys have already been made. It is, however, not altogether easy to decide how far his data really support the inferences that he has drawn. To begin with, he ignores the somewhat special conditions which, according to other social investigators, appear to prevail in Glasgow. Burt, in his early studies at Liverpool, found that the amount of delinquency among children of Irish immigrants or of Catholic parents was disproportionately high; and this was confirmed in a later study by Bagehot. It is conceivable that a similar conclusion might hold good if the Glasgow figures could be analysed according to districts. Professor Ferguson prints figures for 'Church attendance,' but says nothing about religious denomination. And in general "the sociological information which he gives is" (as another critic has remarked) "woefully inadequate."

Medical writers who deal with psychological problems are fond of criticizing "the psychologist's passion for statistical calculations." Accordingly, one of Professor Ferguson's chief aims was "to bring out the relations (between crime and causal factors) without invoking the aid of elaborate statistical techniques." Thus, when a difference is observable in the figures obtained from two contrasted groups, he appears to accept it straight away if it agrees with his preconceptions (even though it is not statistically significant) and to ignore it if it does not. No endeavour is made to use the figures from the larger control group to eliminate the influence of the various conditions as found among the delinquent groups (e.g., by calculating coefficients of contingency or correlation). To demonstrate the effects of subsidiary factors all that is done is to omit those cases in which the major factor is found, and then recalculate the percentages—a somewhat fallacious procedure.

In criticizing the methods adopted by Professor Ferguson in his earlier survey, Mr. Alec Rodger declares that the writer "has slipped on practically every banana-skin there was for him to slip on";¹ and he agrees with the critic in *The Times* who complained that, in their discussions of delinquency, the investigators had constantly confused mere association with actual causation. The present volume endeavours, not (it must be confessed) with very great success, to circumvent these objections. As Mr. Rodger points out, many of the difficulties have arisen because, from the very outset, the essential problems were carelessly formulated in terms very loosely defined, and the whole enquiry as a result was somewhat amateurishly planned. Nevertheless, the tables contain much useful material. There is an obvious moral to be drawn: if those who are experts in medicine wish to launch out on enquiries in the psychological and sociological fields, they should be urged to seek the co-operation, or at least the advice, of those who are familiar with the innumerable pitfalls and trained in the appropriate techniques.

H. L. Jackson

Contributions of Psychology to Social Problems. By Cyril Burt. London: Oxford University Press. 1953. Pp. 76. 5s.

At the beginning of this century, L. T. Hobhouse, the first Professor of Sociology at the London School of Economics and Political Science, was looked upon as the leader in what was then the newly christened field of social psychology. The Webbs, to whom the establishment of the School was chiefly due, had been eager to replace the older, speculative brand of political theory and practice by "a scientific discipline founded on objective methods of research," particularly that of "quantitative study based on first-hand data," which had been introduced by Charles Booth and Francis Galton. "Every conclusion," said Mrs. Webb, "which is derived from observation or experiment, has to be qualified; as well as verified, by relevant statistics."² Accordingly, social psychologists like Hobhouse, Graham Wallas, and the young McDougall—all working together at that time in London—set themselves to formulate, in the shape of tentative hypotheses, the basic problems which appeared to confront the scientific investigator in the new field of work.

¹ *Occupational Psychology*, XXVI, 1952, pp. 251-2.

² Beatrice Webb, *My Apprenticeship*, p. 340.

In his present survey, which forms the substance of his Hobhouse Memorial Lecture, Sir Cyril Burt seeks to summarize the chief advances made during the last fifty years in the many-sided attempts to solve such questions. One of his incidental aims, he tells us, has been to draw attention to British contributions, "partly because recent writers (who appear more familiar with American work) have so often ignored them, partly because it seems rash to assume that conclusions about social matters reached by investigators in other countries will also hold good over here," but above all because he believes that "during and since the recent war British psychologists, while devoting themselves more and more to practical applications, have tended increasingly to neglect research on fundamental issues." He points out that, although research in social psychology has been carried out far more extensively in America than over here particularly during the last two decades, nevertheless, from the time of Booth's surveys onwards, such work has exercised a far more direct effect on legislative and administrative changes in this country than in the United States.

He begins with a brief account of the special scientific techniques—observational, experimental and statistical—devised for such investigations. Most of them, he claims, owe their origin to British pioneers—Galton, Booth, Pearson, the Webbs, and Hobhouse himself. To a large extent, as he points out, they were first tried and refined in the course of investigations on social and educational enquiries conducted with the aid of the schools, where material was more readily accessible.

Turning to the results hitherto attained by the use of such procedures, he considers, in the first part of his review, what are the special mental characteristics of the human race that specially affect social organization and social relations. He concludes that they are chiefly 'similarities in the so-called social instincts' and 'differences in individual ability and temperament.' In the second part he discusses, in somewhat greater detail, the converse problem, namely, the influence of social organization and relations on the mental characteristics of individuals, as they develop in a civilized community. Here he recapitulates the main results achieved by statistical and experimental methods on the problems of sex, social class, and social attitudes and opinions.

The third part summarizes a number of typical enquiries, again based mainly on the combination of case-studies and statistics, in regard to recent trends and changes in religion, morality, leisure occupations, and the like. The whole survey forms a well-packed yet highly readable account of the conclusions so far established, and closes with a strong plea for systematic research in the social branches of psychology comparable to that which has achieved such striking success in the educational branches.

J. S. Mason

The Causes and Treatment of Backwardness. By Cyril Burt. London: National Children's Home, 1952. Pp. 144. 4s. 6d.

This is the most recent of a series of books, published at a surprisingly low price by the National Children's Home, and dealing in non-technical language with problems in child psychology and education. The purpose of the present volume is to explain to the teacher, the parent, and the student of other branches of psychology the need for a scientific investigation of educational sub-lays special stress on the aid given by quantitative procedures, like intelligence-tests and standardized tests for special aptitudes and scholastic attainment, and above all on the importance of a statistical evaluation of the available evidence in order to check impressionistic views about causation and treatment such as general experience or theoretical knowledge may suggest. At the same time he emphasizes still more forcibly his view that tests and statistics by themselves are quite inadequate.

For the practical investigation of individual pupils he proposes a scheme of case-study, and contrasts it with the common type of case-report submitted by teachers and even clinical psychologists, where little or no order is observed either in the investigation or in the written account, and where facts, inferences, personal impressions, and favourite dogmas are all confused together. The character sketches that he himself includes provide vivid descriptions of typical individuals, and demonstrate that it is quite possible to combine human and sympathetic insight with a demand for precise measurement and patient statistical verification. The dry bones live. Incidentally, useful suggestions are made as to the ways in which teachers, particularly those with a professional knowledge of mathematical procedures, might contribute to the advancement of the subject by carrying out researches in their own schools. The practical recommendations on the organization of special schools, special classes, and special remedial schemes suited to each individual case make this work an admirable manual for the beginner in clinical psychology. For the statistical psychologist its chief interest is perhaps to demonstrate how abstract quantitative procedures need not be incompatible with a concrete interest in human personalities.

E. A. Gardner

Facts from Figures. By M. J. Moroney. London : Penguin Books, Ltd. 1951. Pp. viii + 472. 5s.

Mr. Moroney describes his book as "an attempt to take the reader on a conducted tour of the statistician's workshop": the visitor is allowed to watch experts at their task, and then encouraged to try his own hand. Those who are invited are "not only students, but all whose work calls for a general knowledge of the subject, particularly in the world of industry and research." And they are assured that "the rapid expansion and enormous success of statistical techniques in recent years is ample proof of the need for such methods."

The publishers are certainly to be congratulated on producing a book of nearly 500 pages with diagrams, numerous tables, and formulæ that must have cost the compositor some skill—all for the price of two half-crowns. The introductory explanations are racy, lucid, and entertaining. Indeed, every chapter contains ingenious analogies or illustrations which will often prove instructive, not only to the novice, but even to those already acquainted with the stock procedures. Special attention is paid to those branches that are applicable to industrial production and quality control. The more advanced sections are, on the whole, fairly up to date; and the concluding chapters deal with some of the most recent devices in the analysis of covariance, the study of time series, and the use of ranking.

On the other hand, many pages betray signs of haste or carelessness. The theoretical explanations are often inadequate, and some would certainly be frowned upon by statistical critics. There are awkward gaps in the arguments, references to a diagram which does not exist, and numerous misprints, in addition to those contained in the publishers' *errata* (which relate only to four out of the twenty chapters). Some of the distributions are badly grouped, some of the percentages wrongly calculated; and several of the answers to questions given at the end are incorrect. No doubt most of the slips will be corrected in a later edition. Meanwhile, as a semi-popular introduction the book should be of great help to those students who are unable to afford the most recent text-books or who find the more academic expositions of mathematical procedures too dull, too abstract, or positively bewildering.

Cyril Burt

THE RELATION OF THE TERMAN-MERRILL VOCABULARY TEST TO MENTAL AGE IN A SAMPLE OF ENGLISH CHILDREN

By M. I. DUNSDON and J. A. FRASER ROBERTS

Burden Mental Research Department, Stoke Park Colony, Bristol,
and Research Department, Royal Eastern Counties Institution, Colchester

I. Introduction. II. Material. III. The Correlation between Number of Words and Mental Age at Fixed Chronological Age. IV. The Association between Number of Words and Mental Age. V. The Association between Number of Words and Chronological Age. VI. Vocabulary Norms for the Mental Age-Levels of the Terman-Merrill Scale. VII. A Sex Difference. VIII. Summary.

I. INTRODUCTION

The surveys in which we have been interested, including those now in progress, necessitate the testing of large numbers of subjects. For example, we are engaged in a study of low-grade mental deficiency, in which part of the plan is to test the brothers and sisters of the subjects. The most important initial question is whether or not the sibs, other than the few who are also low-grade defectives, are on the average of lower intelligence than the general population: in such work tests are used for comparing relatively large groups, and the interpretation of individual results is not required. The limiting factor is the amount of testing time needed; and the use of individual tests taking an hour or more to administer imposes a severe restriction on what can be done. As an alternative a vocabulary test is an obvious choice. Since Kirkpatrick ((3), 1891; (4), 1907) first suggested a vocabulary test for children, much work has been done on vocabularies, and their merits have been widely recognized. Perhaps the best-known version is that introduced by Terman and Childs ((10), 1912) into their revision of the Binet Scale. Hence it seems unnecessary to do more than summarize their suitability for purposes such as ours, and to indicate briefly the work in progress to which this paper is an introduction.

1. In surveys of this kind, completeness of ascertainment, or something very near it, is essential. There is probably no other type of measurement with which it is easier to get a biased sample than those furnished by intelligence-test scores. Outside the limits of school age it is almost impossible to secure a complete sample, so that the age of the subjects must be limited to 5-15 years. But in a survey such as that mentioned all subjects within those limits of age are needed, for the amount of information increases rapidly with the size of the sibship. Hence group-tests are of no use. Vocabulary provides an individual test which is essentially the same test over the whole of the required span of age.

2. It is known that the reliability of vocabulary tests is in general quite high. As a rule it appears to average about .90.

3. Validity too is very high. In the sample described in this paper the partial correlation between number of words and mental age at fixed chronological age is .84, a figure very similar to those found by previous investigators.

4. It seems likely that practice effect is minimal. The effect of mere repetition of the same vocabulary is apparently nil. McMeeken ((5), 1939) describes a retest with the Stanford (1916) Scale

of 140 children after the lapse of one day. The mean increase of I.Q. was 2.46 points; but she states that in vocabulary score no change was observed. We are not aware of any work on the effect of attempts to coach children in the general task of word definition; but it seems unlikely that this could have any appreciable success. The only factor likely to be serious would be practice in defining the actual words to be used. This can be obviated by providing alternative vocabularies, though of course much heavy work is needed to standardize a new one.

5. Above all, there is the great saving in testing time. A partial correlation of .84 between M.A. and words defined at fixed C.A. means that 70 per cent. of the information is being obtained in 10 per cent. of the testing time. With a test as applied to an individual the nature of the 70 per cent. and of the residual 30 per cent. raises important problems; but for the comparison of large groups this need not concern us greatly, though naturally one would be careful to consider the possible associations of so verbal a test with such factors as home background and social class.

With these considerations in mind, one of us (M. I. D.) is administering four vocabularies to a relatively large number of children in the City and County of Bristol, the children being selected by a method which, it is hoped, will yield a good random sample. The four vocabularies are the Terman-Merrill, the Mill Hill Vocabularies A and B, and the vocabulary from the Wechsler Intelligence Scale for Children. We are grateful to Professor L. M. Terman, Mr. J. C. Raven, and Dr. David Wechsler for giving permission for their use.

We have, of course, no intention of standardizing the vocabularies in terms of anything other than themselves, and entirely agree with McNemar ((6), 1942) about the impropriety of attempting to convert vocabulary score into some kind of equivalent in mental age, and thereafter working out an I.Q. Nevertheless, it is valuable to be able to relate one of the vocabularies to a complete scale. We were also anxious to know how far the high average correlation taken over a wide span of age varied at different age-levels and whether it is justifiable to use a vocabulary as a single test down to the age of 5 years. Should there be a sex difference, as indeed we have found in our sample, it is necessary to know whether and to what extent mental age is also involved. Vernon ((12), 1949) has published results based on a sample of English children, showing that their vocabulary scores at most of the relevant mental ages considerably exceeded those required for a pass. For these reasons, and because we had a sample already available, we thought that it might be useful to put on record data for yet another sample of English children.

II. MATERIAL

The main part of the data consists of the results obtained by applying the Terman-Merrill test (Form L) to 450 children ranging in age from 7.0 to 14.11 years. The material was assembled for another purpose in connexion with a recent paper (Dunsdon (2), 1953). The tests were given by Mrs. M. D. Jenner and Mrs. A. E. Sedgley, working at the Burden Mental Research Department of Stoke Park Colony, and by Dr. Margaret A. Mellone and Mrs. E. H. Scarr, working at the Research Department of the Royal Eastern Counties Institution, Colchester. They were applied for a variety of purposes. Many of the children were sibs of mental defectives tested in a survey of which some preliminary account has already been published (Roberts (8), 1952); but there were a number of other smaller groups. It was decided to make up a sample with a mean I.Q. of 103 (probably closer to the population mean than 100) and a standard deviation of 17. The limiting factor was the number of persons in the higher (though not the highest) ranges for whom test-results were available. In that range of I.Q. in which there were proportionately fewest results all were used. At other parts of the range the appropriate number was selected by a random process from those available. For the present paper 52 records were added for children of C.A. 5.0-6.11, the mean I.Q. and standard deviation being respectively 96.3 and 15.3. Although no attempt was made to select in relation to other factors, such as sex or chronological age, the sample is satisfactory as far as has been discovered except in one respect. A number of the records showing the highest I.Q.s were drawn from those of children (mostly in their twelfth year) who had secured free places at a grammar school; hence the children in the age-group 11.0-11.11 are of relatively higher intelligence than those in other parts of the range. The effect of this is mentioned later.

III. THE CORRELATION BETWEEN NUMBER OF WORDS AND MENTAL AGE AT FIXED CHRONOLOGICAL AGE

Using the smallest possible units of grouping, months and single words, the over-all correlation coefficients are as follows: C.A. and M.A., .8000; C.A. and words, .7349; M.A. and words, .9307. The last figure is very close to Raven's ((7), 1948) correlation of .926 between Mill Hill Vocabulary score and Terman-Merrill M.A. The relevant figure is the partial correlation of M.A. and number of words at fixed C.A. The influence of chronological age can be appropriately removed by the use of partial correlation as it is a measurement not subject to error. It is true that the regressions are not linear, but as the linear functions account for the great bulk of the associations it is not misleading to use them. Furthermore, the non-linear components in the regressions involving C.A. are almost entirely due to the 11-year age-group containing too many children of high intelligence, as mentioned in the previous section. As regards the association of M.A. with C.A., the expected falling off in the rate of increase of M.A. after the age of 12 is shown by our sample, but is not significant with these numbers. The partial correlation is .842. A part of the association is spurious because vocabulary is used in scoring M.A.; but this part cannot be large, as the contribution of vocabulary over the relevant mental ages does not average more than 10 per cent.

The correlations between M.A. and number of words, calculated separately for each year group of C.A., are shown in Table I.

TABLE I. CORRELATION BETWEEN MENTAL AGE AND VOCABULARY AT SUCCESSIVE CHRONOLOGICAL AGES

Chronological Age	Number	Correlation
5.0-5.11	18	.89
6.0-6.11	34	.71
7.0-7.11	40	.60
8.0-8.11	46	.72
9.0-9.11	54	.82
10.0-10.11	67	.86
11.0-11.11	67	.83
12.0-12.11	54	.90
13.0-13.11	65	.85
14.0-14.11	57	.89

The figures are closely similar to those of other workers. Terman and Merrill ((11), 1937) state that the correlations between vocabulary and the scale as a whole for single age-groups range from .65 to .91 with an average of .81. McNemar ((6), 1942) gives correlations with the composite scale of .71, .83, .86, and .83 at ages 8, 11, 14, and 18 respectively.

IV. THE ASSOCIATION BETWEEN NUMBER OF WORDS AND MENTAL AGE

The association of M.A. and number of words can now be examined in more detail, the basic data being shown in Table II. The relation is clearly non-linear. Perhaps some apology is needed for curve-fitting; but it seems to us that it is helpful to obtain smoothed curves by efficient methods. When the fit proves to be very good, as in this instance, with

TABLE II. BIVARIATE FREQUENCY TABLE FOR MENTAL AGE AND NUMBER OF WORDS

Mental Age	Vocabulary : Number of Words																													Total
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
19.0-19.11																													1	1
18.0-18.11																													1	1
17.0-17.11																													2	1
16.0-16.11																													1	1
15.0-15.11																													3	1
14.0-14.11																													2	1
13.0-13.11																													4	1
12.0-12.11																													5	1
11.0-11.11																													4	1
10.0-10.11																													2	1
9.0-9.11																													1	1
8.0-8.11																													4	1
7.0-7.11																													3	1
6.0-6.11																													2	1
5.0-5.11																													1	1
4.0-4.11																													1	1
3.0-3.11																													1	1
Total	2	1	4	3	2	13	27	20	53	48	34	35	23	28	29	22	27	23	22	12	13	13	11	14	8	6	2	1	5	1
																														502

entirely non-significant remainders which are close to expectation, then one can feel that the data have been reduced to their simplest form, a form which permits trends to be perceived, unobscured by chance fluctuations. It turns out that biquadratics are needed for both regressions, the equations being:

$$y = -12.64 + 6.941x - 0.9849x^2 + 0.06982x^3 - 0.001664x^4,$$

where y = number of words and x = M.A. ($3.0 - 3.11 = 3$, etc., . . .):

$$x = 2.882 + 0.2696y + 0.2675y^2 - 0.02691y^3 + 0.0008121y^4,$$

where x = M.A. as before and y = number of words in units of grouping ($0 - 1 = 1$, $2 - 3 = 2$, etc., . . .).

The fitting is shown in Table III and the curves are shown graphically¹ in Fig. 1. The

TABLE III. RELATION OF MENTAL AGE AND VOCABULARY: ANALYSIS OF VARIANCE

A. Regression of Vocabulary on Mental Age

Variation of Vocabulary	Degrees of Freedom	Sum of Squares (words) ²	Mean Square (words) ²	Variance Ratio	Significance
Linear regression	1	13,771.94	13,771.94	35.19	++
Addition of quadratic term	1	151.95	151.95	0.88	+
Addition of cubic term	1	3.79	3.79	13.85	++
Addition of biquadratic term	1	58.32	58.32	1.13	-
Remainder between arrays	12	57.02	4.75		
Within arrays	485	2,035.47	4.20		
Total	501	16,078.49	32.09		

B. Regression of Mental Age on Vocabulary

Variation of Mental Age	Degrees of Freedom	Sum of Squares (years) ²	Mean Square (years) ²	Variance Ratio	Significance
Linear regression	1	4,356.17	4,356.17	44.82	++
Addition of quadratic term	1	63.64	63.64	0.22	+
Addition of cubic term	1	0.32	0.32	5.29	+
Addition of biquadratic term	1	7.46	7.46	0.91	-
Remainder between arrays	10	12.86	1.29		
Within arrays	487	688.09	1.41		
Total	501	5,128.54	10.24		

quadratic terms are relatively large in both instances. The biquadratic terms effect a substantial improvement; with the regression of words on M.A. the residual variance between arrays is cut down to a half and with the other regression by a third. The form of the curves throws light on the age-group correlations of Table I. The most important feature is the steep rise at the youngest ages, followed by a relative flattening and then by a steep rise once again. The high correlation at age 5, the fall to a minimum at age 7, and then the rise almost to the maximum by age 9 are just what would be anticipated.

¹ We are indebted to Mrs. M. G. Young for drawing the figure.

The Relation of the Terman-Merrill Vocabulary Test to Mental Age

It is not a matter apparently of vocabulary being a less meaningful test for the youngest children, either necessarily so, or because the lower ranges of the scale are less verbal. Rather, it would seem, among young children the task is such that vocabulary correlates very highly with other mental measurements; then there is an interval during which the number of words defined becomes rather less meaningful; finally, by age 9 or 10 vocabulary has regained its close association with other tasks.

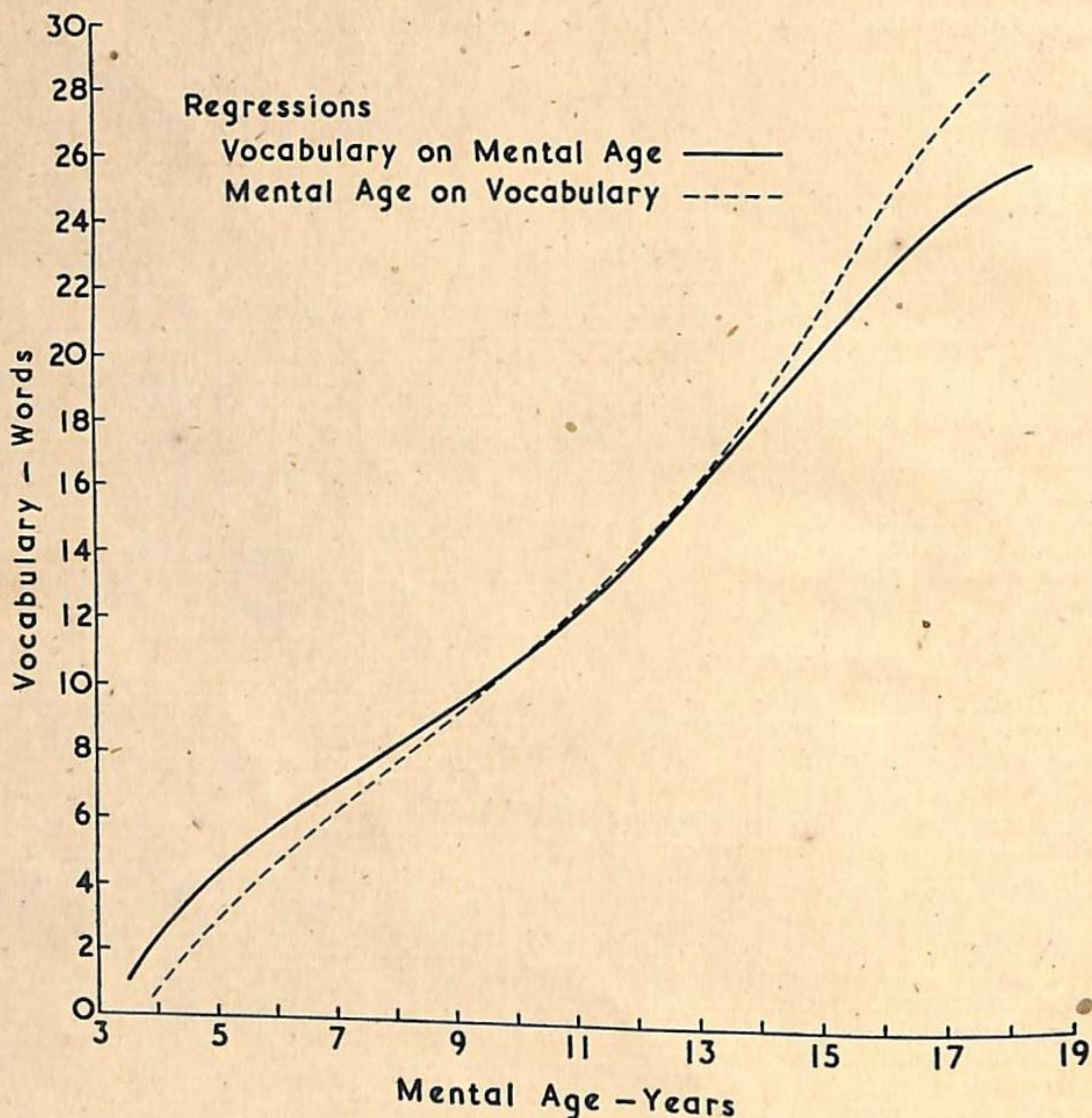


FIG. 1.—Regression of number of words on mental age and regression of mental age on number of words.

It must be recognized, therefore, that from about age 6 to about 8 or 9 validity in terms of the Terman-Merrill Scale is reduced, though not to such an extent that from the practical point of view it is unreasonable to use vocabulary as a single test over the whole range of 5 to 15 years.

V. THE ASSOCIATION BETWEEN NUMBER OF WORDS AND CHRONOLOGICAL AGE

Table IV gives the bivariate frequency table for number of words and chronological age. The association between C.A. and vocabulary is so much looser than that between M.A.

TABLE IV. BIVARIATE FREQUENCY TABLE FOR CHRONOLOGICAL AGE AND NUMBER OF WORDS

Chrono- logical Age	Vocabulary : Number of Words																													Total	
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28		29
14.0-14.11								1	1			3	2		6	4	7	6	5		1	3	3	4	5	1	1	3	1	57	
13.0-13.11								1			1	1	1	7	6	5	4	4	10	7	4	1	4	3	1	2	2		1	65	
12.0-12.11										5	4	4	1	5	4	3	8	4	1	2	3	2	2	4	1	1			1	54	
11.0-11.11									1	2	6	8	3	4	6	2	5	6	4	2	4	7	2	2	1	2				67	
10.0-10.11						1	1	2	7	11	4	8	10	6	4	3	3	2	2	1	1				1					67	
9.0- 9.11								2	12	9	6	6	4	5	1	3		1						1						54	
8.0- 8.11								4	13	10	7	3	2	1	2															46	
7.0- 7.11						4	3	5	12	9	5	2																		40	
6.0- 6.11	1	2	2	1	1	5	11	6	4	1	1																			34	
5.0- 5.11	2	2	1	1	3	4	4	1	3	1																				18	
Total	2	1	4	3	2	13	27	20	53	48	34	35	23	28	29	22	27	23	22	12	13	13	11	14	8	6	2	1	5	1	502

and vocabulary that with our numbers there is not much point in examining the form of the regression lines, particularly since in this sample, as already mentioned, the great bulk of the departure from linearity is due to the accidental circumstance that too many clever 11-year-old children were included. More ample data from a better sample will be forthcoming when the main enquiry is completed. The difference in appearance between the two bivariate frequency tables (Tables II and IV) is, however, very striking. The growth in variability of number of words seems largely to keep pace with the growth in the variability of M.A., but in terms of C.A. the variability of the number of words defined increases very markedly with age. It will be necessary to allow for this in the age standardization of the vocabularies.

VI. VOCABULARY NORMS FOR THE MENTAL AGE-LEVELS OF THE TERMAN-MERRILL SCALE

As regards total score there is reasonably good agreement between the norms given by Terman and Merrill and the results obtained with British children. Nevertheless, many individual items do show differences, some being placed too high for British children and some too low (Burt (1), 1948). The standardization of the vocabulary test has been criticized by several writers as being much too easy. Vernon ((12), 1949, p. 184) gives a table of norms, based on a preliminary research, which differ widely from those given in *Measuring Intelligence*. Table V compares the norms derived from the present sample with those of Terman and Merrill and of Vernon.

TABLE V. NUMBER OF WORDS CORRECTLY DEFINED AT MENTAL AGE-LEVELS IN TERMAN-MERRILL SCALE

Mental Age	No. of Words Correctly Defined		
	Terman and Merrill's Sample	Vernon's Sample	Present Sample
6..	5	5	6
8..	8	9	8
10..	11	13	11
12..	14	17	14
14..	16	21	18
15.4 (A.A.)	20	24	21
17.4 (S.A. I)	23	29	25
19.10 (S.A. II)	26	36	—
22.10 (S.A. III)	30	42	—

The expectations based on the present sample are identical with those of Terman and Merrill at M.A. levels 8, 10, and 12. The difference of one word at M.A. 6 hardly means anything as our expectation is only just over 5.5. At M.A. levels 14, A.A., and S.A. I, our figures lie between those of the other series, but are closer to those of Terman and Merrill. It would be unjustifiable with our sample to extrapolate to S.A. II, but our curve suggests a number much nearer 26 than Vernon's 36. It should be mentioned, however, that it is possible that the present sample may be somewhat overweighted with children from lower social groups; and if this is so it may be that the number of words defined has been slightly reduced in relation to the scale as a whole.

VII. A SEX DIFFERENCE

The present sample shows a small but highly significant sex difference. The means and linear regressions are as follows :

	No.	Mean C.A. (years)	Mean M.A. (years)	Mean No. of words	Increment in words per year of C.A.	Increment in words per year of M.A.
Boys ..	260	10.635	10.881	13.150	1.674	1.681
Girls ..	242	10.706	10.940	12.665	1.547	1.627
Difference ..		-0.071	-0.059	+0.485	$\pm 0.127 \pm 0.133$	$\pm 0.054 \pm 0.058$

Using the regressions to adjust M.A. and C.A. to 10.80 years, the sex differences in favour of boys in number of words defined as at that age are : C.A., 0.616 ± 0.343 ; M.A., 0.575 ± 0.186 . Thus, for a given C.A. or M.A., boys define more words than girls, the difference as at 10.80 years being about 0.6 word. The difference is not significant for C.A., but, owing to the close association of number of words and M.A., is highly significant for M.A., being 3.1 times its standard error. It would seem to extend over the whole range of age, and is apparently a matter of the position of the regression lines and not of their slope, though it is possible that the rates of growth might be significantly dissimilar with larger numbers.

It will be noted that it is vocabulary which is slightly sex biased and not mental age as a whole. In another British sample, that of the Scottish Survey (7), 1949, boys did better on the Terman-Merrill Scale than girls ; but that sample covered only a narrow range of age. In the present sample, with its 10-year span, there is no trace of sex bias with the scale as a whole. Girls are 0.071 year older than boys and their mean M.A. is 0.059 year greater, a difference of one in the first decimal place of I.Q. Agreement between the sexes could hardly be closer.

VIII. SUMMARY

1. Intelligence-tests intended to cover the whole range of school age must be applied individually. For the comparison of large groups vocabulary tests have manifest advantages in respect of reliability, validity, relative absence of practice effect, and, above all, in the amount of information obtained per unit of testing time. Accordingly four vocabularies are being given to school children at Bristol, selected by a method which, it is hoped, will give a good random sample. The present paper presents data from a sample of English children showing the relation of the Terman-Merrill vocabulary test to the scale as a whole.

2. The sample consists of results obtained by applying Form L to 450 children ranging in age from 7.0 to 14.11 years. These were selected from the results available so as to yield a sample with a mean I.Q. of 103 and a standard deviation of 17. To them were added data for 52 children aged 5.0 to 6.11 years, who were slightly lower in I.Q. and slightly less variable.

3. The over-all partial correlation between number of words and M.A. at fixed C.A. is .842, which is in close agreement with the results of other workers. Thus 70 per cent. of the information is obtained in 10 per cent. of the testing time.

4. The regressions of words on M.A. and of M.A. on words are non-linear. The form of the curves is reflected in lower correlations between number of words and M.A. in the seventh, eighth, and ninth years of life, these being respectively .71, .60, and .72. The other age-group correlations range from .82 to .90.

5. The number of words defined at the various M.A. levels of the scale agree almost exactly with those of Terman and Merrill up to 12 years. At older levels the children of the present sample exceed the norms given in *Measuring Intelligence*, but are closer to these norms than to those given by Vernon.

6. In this sample there is no sex difference in performance with the Terman-Merrill Scale as a whole. There is, however, a sex bias in vocabulary score ; for given M.A. boys define more words than girls.

REFERENCES

1. Burt, C. (1948). *Handbook of Tests for Use in Schools* (2nd ed.). London : Staples Press.
2. Dunsdon, M. I. (1953). 'A Comparison of Terman-Merrill Scale Test Responses among Large Samples of Normal, Maladjusted, and Backward Children.' *Jour. Ment. Sci.* (in the press).
3. Kirkpatrick, E. (1891). 'Number of Words in an Ordinary Vocabulary.' *Science*, XVIII, 107-8.
4. Kirkpatrick, E. (1907). 'A Vocabulary Test.' *Pop. Sci. Monthly*, LXX, 157-64.
5. McMeeken, A. M. (1939). *The Intelligence of a Representative Group of Scottish Children*. London : University of London Press.
6. McNemar, Q. (1942). *The Revision of the Stanford-Binet Scale*. Boston : Houghton Mifflin.
7. Raven, J. C. (1948). 'The Comparative Assessment of Intellectual Ability.' *Brit. J. Psych., Gen. Sect.*, XXXIX, 12-19.
8. Roberts, J. A. F. (1952). 'The Genetics of Mental Deficiency.' *Eugen. Rev.*, XLIV, 71-83.
9. Scottish Council for Research in Education (1949). *The Trend of Scottish Intelligence*. London : University of London Press.
10. Terman, L. M., and Childs, H. G. (1912). 'A Tentative Revision and Extension of the Binet-Simon Scale.' *J. Educ. Psychol.*, III, 205-8.
11. Terman, L. M., and Merrill, M. A. (1937). *Measuring Intelligence*. London : Harrap.
12. Vernon, P. E. (1949). *The Measurement of Abilities*. London : University of London Press.

CHANGES IN Terman-Merrill I.Q.s WITH DULL CHILDREN

A Test of the Roberts-Mellone Adjustments

By ELIZABETH H. SCARR

Research Department, Royal Eastern Counties Institution, Colchester

I. *Introduction.* II. *The Data.* III. *Comparison of I.Q.s in Test and Retest.*
IV. *Comparison between the Observed Mean Changes in I.Q. and those Predicted*
by the Roberts-Mellone Adjustments. V. *Summary.*

I. INTRODUCTION

In a recent paper¹ Roberts and Mellone have devised a method of adjusting Terman-Merrill I.Q.s for their lack of independence of chronological age. They describe the fitting of a curve for regression of variance of I.Q. on C.A., obtained by using the test results (Form L) of the 2,000 children in the Terman-Merrill sample whose ages range from 5.0 to 14.11 years. From this curve is derived a table of adjustments, so that the adjusted I.Q.s become comparable at different age levels. In their paper Roberts and Mellone say:

"An I.Q. of 52 at 5 years 1 month may be expected to rise to a maximum of 54 at 5 years 7 months; remain constant for about a year; fall by about 2 points a year to age 8; thereafter fall steadily at the rate of 4 points a year to 11 years 1 month; remain constant to 13 years 1 month; and finally rise to 49 again. From 14 years 7 months or so the rise becomes suspiciously great. Whether in fact these actual changes will be shown on large samples can only be determined by repeated measurements of the same individuals."

Data for making this practical test have since become available in connexion with retests made on a sample of mentally defective and backward children as part of a study on mental growth at low intelligence levels. As the adjustments are based on variances, their average size in any given group increases with the distance of that group from the mean of the general population. In a group of relatively low I.Q. the adjustments are large; and consequently the sample thus examined provides suitable material for making the test.

II. THE DATA

The sample consists of a group of backward and mentally defective children attending Special Schools for the educationally subnormal and Occupation Centres in Chelmsford, Colchester, Dagenham, Leytonstone, and Walthamstow, and the Residential Special Schools of the Royal Eastern Counties Institution in Colchester and Cambridge. All the available children in the selected schools, between the ages of 6.0 and 14.11 years, were tested by means of Form L. (Children under the age of 6 years were not included, partly because relatively few find their way into Special Schools and Occupation Centres, and partly because very retarded children are untestable when so young.) So far as possible, the same children were retested, after the lapse of one year. Some difficulty was encountered in obtaining a retest after this fixed

¹ This *Journal*, V (1952), pp. 65-79.

interval, the main obstacles being absence through illness and the intervention of school holidays. About 10 per cent. of the original group could not be retested on a calendar date sufficiently close to that of the original test; and these were omitted. The complete sample consists of 402 subjects, the age range at first test being 6.0 to 14.11 years. All tests and retests were made by myself.

For testing the adjustments only those children aged 13.11 years or less at the first test could be used, since the correction table for Terman-Merrill I.Q.s gives adjustments to age 14.11 years only. This smaller group consists of 350 subjects¹ whose ages range from 6.0 to 13.11 years at first test. The I.Q.s range from 20 to 85 at the first test, with a mean of 56.97 and a standard deviation of 14.12.

The intervals between first and second tests varied about 365 days in accordance with the following distribution: $\pm 0-7$ days, 207 children; $\pm 8-14$ days, 79 children; $\pm 15-21$ days, 41 children; $\pm 22-28$ days, 23 children.

III. COMPARISON OF I.Q.s IN TEST AND RETEST

TABLE I. FREQUENCY DISTRIBUTION OF DIFFERENCES IN I.Q. BETWEEN FIRST AND SECOND TESTS

2nd I.Q. minus 1st I.Q.	1st I.Q. < 40	1st I.Q. > 39			Total
	All C.A.s	C.A. at first test			
		6.0-10.11	11.0-11.11	12.0-13.11	
+ 10				2	2
+ 9				2	3
+ 8				2	2
+ 7				2	1
+ 6		1			5
+ 5	2	3	1	1	8
+ 4	3	4		2	16
+ 3	3	6	1	6	20
+ 2	1	5		12	21
+ 1	7	8	1	11	27
0	12	9	3	8	35
- 1	8	9	6	8	35
- 2	5	9	3	15	38
- 3	7	16	5	12	39
- 4	2	18	4	10	25
- 5	3	10	4	9	23
- 6	2	9	4	7	17
- 7	1	10	3	2	10
- 8		5	2	2	11
- 9		7	2	2	5
- 10		3	1	1	3
- 11		1	1		
- 12					
- 13		1	1		2
- 14		1			1
		1			1
Number	57	136	42	115	350
Mean	- 0.561	- 2.191	- 2.976	- 0.357	- 1.417
Standard Error of Mean	0.403	0.355	0.575	0.367	0.213

¹ The actual number of tests

¹ The actual number of children was 339, of whom 11 were tested three times. The 11 are too few for separate treatment and the analysis is made in terms of the 350 one-year differences.

Table I shows the frequency distributions of differences between test and retest. (a) The mean difference between the two tests, for the whole group of 350 subjects, is -1.417 ± 0.213 ; the standard deviation of differences is 3.977. (b) The mean difference for the whole group, ignoring sign, is 3.320 ± 0.139 , with a standard deviation of 2.604. (c) The coefficient of correlation between first and second I.Q. is $+0.960$.

The grouping of chronological age and I.Q. used in Table I anticipates the findings of the next section. It will be seen that this grouping shows most clearly the trends of agreement and disagreement between the observed changes in I.Q. and those predicted by the Roberts-Mellone adjustments.

IV. COMPARISON BETWEEN THE OBSERVED MEAN CHANGES IN I.Q. AND THOSE PREDICTED BY THE ROBERTS-MELLONE ADJUSTMENTS

The method of finding the predicted change in I.Q. during the year between the two tests was as follows. The I.Q. obtained in Test 1 was located in the correction table, Table VIII (*loc. cit.*, pp. 72-3), opposite chronological age for Test 1, and the adjustment at the head of the column noted. Then the same obtained I.Q. at Test 2 was again located in the body of the table, this time opposite chronological age for Test 2 (i.e., one year later), and again the appropriate adjustment at the head of the column was noted. The difference between these two adjustments gives the predicted rise or fall in I.Q. over the period of the one year between the two tests. For example, for I.Q. 63 at age 10.0 years, the correction is $+5$; for the same I.Q. at age 11.0 years, it is $+7$. Thus the predicted fall in I.Q. over the year is 2 points. We have, then, for each subject the observed change in I.Q. and the predicted change in I.Q. By comparison of the two the applicability of the adjustments can be tested at the various age and I.Q. levels of the 350 subjects.

The observed and predicted mean changes in I.Q. are shown in Table II.

Theoretical considerations as well as practical experience indicate that there must be a point in intelligence-level, though not necessarily a clearly defined one, below which adjustments are inapplicable. The figures of Table II show that in the present sample an I.Q. of about 40 hits off this point very fairly. In comparing the observed and predicted mean changes, therefore, I.Q.s under 40 have been omitted from the means shown in the right-hand totals column of Table II. This also explains why in Table I a separate frequency distribution of differences is shown for all I.Q.s under 40.

From inspection of Table II it can be seen that the agreement between observed and predicted changes in I.Q. is, on the whole, very good up to the age of 10.11 years at first test. As the predicted change in a child of 10.11 years refers to growth from that age up to 11.11 years, this means that the adjustments are applicable up to a chronological age of about 12 years. For this span of chronological age, 6 to 12 years, the Roberts-Mellone fitted curve for variance of I.Q. on C.A. predicts a substantial fall in I.Q. over the lower ranges of I.Q. This is clearly borne out by the data. During the thirteenth year, the curve predicts a period of stability, during which little change in I.Q. takes place; the actual results, however, show a continued fall in I.Q. During the fourteenth and fifteenth years, the Roberts-Mellone curve predicts a fairly steep rise in low I.Q. (As the variance decreases, low I.Q.s should be rising.) Instead of this, the data show that the levelling off, which was predicted in the thirteenth year, is in fact taking place in the fourteenth and fifteenth years.

TABLE III. SUMMARY OF OBSERVED AND PREDICTED MEAN CHANGES IN I.Q.

I.Q. at 1st test	C.A. at 1st test								
	6.0-10.11			11.0-11.11			12.0-13.11		
	No.	Obs.	Pred.	No.	Obs.	Pred.	No.	Obs.	Pred.
40-49	16	- 1.8	- 2.6	3	- 3.7	- 0.3	17	- 0.2	+ 3.2
50-59	35	- 0.9	- 2.5	16	- 1.9	- 0.1	32	+ 1.0	+ 3.0
60-69	48	- 2.7	- 1.9	12	- 3.6	- 0.3	45	- 0.7	+ 2.2
> 69	37	- 2.9	- 1.7	11	- 3.6	- 0.3	21	- 1.8	+ 1.4
Total	136	- 2.19	- 2.08	42	- 2.98	- 0.21	115	- 0.36	+ 2.44
Std. Error of Obsd. Mean		± 0.355			± 0.575			± 0.367	

Table III, which is a summary of the results shown in Table II, clearly shows these trends.

In their paper, Roberts and Mellone also describe corresponding calculations carried out on test results from a sample of 689 children between the ages of 5.0 and 14.11 years, being the sibs and cousins of a group of mentally defective children. This sample was obtained from areas around Bristol and Colchester. The correspondence with the Terman-Merrill sample was close except in the oldest children. An alternative series of multipliers, based on the English sample, was included for comparison in their Table VIII. It seemed worth while calculating the changes of I.Q. predicted on the basis of this latter sample. It might have been thought that the observed changes in I.Q. would correspond more closely with those predicted on the basis of a sample of English children. This was not, however, found to be so, as is clearly revealed in Table IV. With the solitary exception of ages 9.0 to 9.11 years, there is closer agreement with the changes predicted on the basis of the Terman-Merrill sample.

TABLE IV. COMPARISON OF OBSERVED MEAN CHANGES WITH THOSE PREDICTED ON THE BASIS OF TERMAN-MERRILL AND ENGLISH SAMPLES

C.A. at 1st test	Number	Mean change		
		Observed	Predicted	
			Terman-Merrill Sample	English Sample
6.0- 6.11	8	- 1.75	- 2.25	- 3.13
7.0- 7.11	16	- 0.94	- 2.81	- 3.44
8.0- 8.11	29	- 1.97	- 2.66	- 3.00
9.0- 9.11	39	- 2.51	- 2.05	- 2.13
10.0-10.11	44	- 2.59	- 1.47	- 1.07
11.0-11.11	42	- 2.98	- 0.21	+ 0.26
12.0-12.11	58	- 0.67	+ 1.09	+ 2.36
13.0-13.11	57	- 0.04	+ 3.82	+ 8.04

There remains the question of the effect of practice in the test on the second obtained I.Q. As the interval between Test 1 and Test 2 was one year, and the group consisted of children of subnormal intelligence, it has been assumed that the practice effect was negligible. If it had been appreciable, one would have expected it to manifest itself in the higher I.Q. ranges. The results, however, show no evidence for this.

V. SUMMARY

1. In a recent paper Roberts and Mellone (1952) examined the relation of variance of Terman-Merrill I.Q. to C.A. and devised adjustments intended to correct I.Q.s for their lack of independence of chronological age.

2. The retesting of 350 mentally defective and backward children after the lapse of a year enables a test to be made of the validity of the adjustments. The children ranged in age at first test from 6.0 to 13.11 years, with a mean I.Q. of 57.

3. In children of all ages the results show that the adjustments are inapplicable to I.Q.s below a level of about 40.

4. With I.Q.s of 40 and over the changes predicted by the adjustments are closely realized up to the age of about 12 years.

5. During the thirteenth year the adjustments presuppose a levelling off, whereas in the sample the mean I.Q. continues to fall. During the fourteenth and fifteenth years the adjustments presuppose a rise in I.Q., whereas in the sample there is a levelling off. Thus in this sample, apparently, the curve of variance of I.Q. on C.A. is inflected about a year later than was predicted.¹

¹ *Acknowledgements.* I am greatly indebted to the Governors of the Royal Eastern Counties Special Schools for the facilities provided and to the head teachers and staff of the schools for their help. I am also greatly indebted to the education authorities and to the head teachers and staff of the Special Schools at Chelmsford, Colchester, Dagenham, Leytonstone, and Walthamstow, and to Dr. Kenneth Cowan and the Supervisors of the Junior Occupation Centres of the County of Essex. My special thanks for help and advice are due to Dr. R. M. Bates, Dr. J. A. Fraser Roberts, and Dr. Margaret A. Mellone.

THE SUBMISSIVENESS OF NATIONS

By LEWIS F. RICHARDSON

Hillside House, Kilmun, Dunoon, Argyll, Britain

I. *Introduction.* II. *The Linear Theory of Arms-races.* III. *First Hypothesis about Submissiveness.* IV. *Second Hypothesis about Submissiveness.* V. *Third Hypothesis (for Sides Equal in their Constants).* VI. *Conclusion.*

I. INTRODUCTION

The problem to be discussed in this paper can be stated quite simply as follows. It is well known that a thorough military defeat leaves the vanquished nation for some years in a mood of submission to its victors (examples are the submissive behaviour of Vichy-France to the Germans during 1941 to 1943, or of West Germany to the Western Powers and of East Germany to the Soviet Union during 1945 to 1952). As this effect seems universal, it is not necessary to name the nations; they may be distinguished as A and B. Now suppose that A and B are both rested from war and are roughly equal in size and in industrial power; and suppose further that A makes a show of force against B. Does such behaviour induce in B a submissive mood or a retaliatory mood?

Whatever may be thought of the following treatment, there can be no doubt that some of the questions connected with submissiveness are important. For example, in September, 1952, the North Atlantic Treaty Organization held combined naval manœuvres (exercise "Mainbrace") off Norway and in the Baltic simultaneously with land manœuvres (exercise "Holdfast") involving nearly 200,000 troops in Western Europe. The declared intention of N.A.T.O. was to deter any tendency in the Soviet Union towards aggression. On the contrary the Russian newspaper *Pravda* said that exercise "Mainbrace" was itself aggressive (B.B.C., September 13th). Personally I accept N.A.T.O.'s description of its intention; but I appreciate that *Pravda* may find this intention very difficult to believe, because of the universal tendency to suspect foreign forces unless those of allies—a tendency represented in the following formulæ by the defence coefficients k or l . To be deterred from any tendency towards aggression is here regarded as a mild form of submission.

Professor C. A. Mace asked whether my paper contained any suggestion for research on small groups of people, a type of enquiry now being pursued by many psychologists. Probably it does; for submission has been occasionally observed wherever two or more individuals meet. For examples: even between wolves (Lorenz (14), 1952, ch. 12); or the pecking order in the hen yard (references are given by Landau (11), 1951); or tests on individual persons in their relation to the community (Allport (1), 1945, pp. 410–14). W. McDougall (1917 (15), pp. 62–6) took subjection to be one of the primary human instincts. It might be rash, however, to expect that by research on a group of, say, 50 people one could predict the behaviour of a nation: for, at least in physics, large aggregates of molecules show some conspicuous forms of behaviour that are not noticeable in small aggregates. For example, I doubt whether anybody, by intensive observation of a drop of water, could have discovered the peculiar law of eddy-diffusion which operates in the sea (Richardson and Stommel (27), 1948).

II. THE LINEAR THEORY OF ARMS-RACES

This theory must first be mentioned, because submissiveness was formulated by a modification of it. From a pamphlet by Bertrand Russell (28) in 1914 I took the idea that much of what is nowadays called defence does not in fact defend, because between

equal groups in peacetime it provokes retaliation. Gregory Bateson ((2), 1936, p. 266) observed competitive boasting in New Guinea and generalized the concept as 'symmetrical schismogenesis,' of which an arms-race is another example. Russell's concept of mutual stimulation was formulated by me as a pair of simultaneous differential equations (1919, revised 1935). In 1938 I hit on a way of testing the formulae quantitatively in terms of expenditure. The testing was continued in a monograph published a few months before the war (*Generalized Foreign Politics*, *Brit. J. Psychol. Mon. Sup.*, 1939; a better title would have been *The Theory of Arms-races*).

The financial facts have been re-examined in much greater detail in a microfilm called *Arms and Insecurity* ((23), 1949), where the technical term 'warfinpersal' is introduced as a contraction for 'war finance per salary.' The 'warfinpersal' of any nation is its governmental warlike (alias 'defence') expenditure divided by the annual earnings of an average sort of citizen, usually those of a semi-skilled engineer. The latest summary, including a glance at the present arms-race, is being printed for *Sankya* (26). The sceptical reader is referred to the aforesaid publications for the facts; here there is room only for the equations which summarize them. As we pass from the study of individual-differences to that of nations, the vagaries of individuals merge into the quasi mechanical regularities of mass-behaviour.

Let x and y denote the warfinpersals of the opposing sides, t the time, and $g, h, \alpha, \beta, k, l$ be constants of which α, β, k, l are positive. The hypothesis is that

$$dx/dt = g - \alpha x + ky, \quad (2/1)$$

$$dy/dt = h - \beta y + lx. \quad (2/2)$$

In particular, if $g = 0, h = 0, k = l$, and $\alpha = \beta$, then the solution is

$$x + y = Ae^{(k-\alpha)t}, \quad x - y = Be^{-(k+\alpha)t}, \quad (2/3, 2/4)$$

where A and B are arbitrary constants. Functions like $x + y$ and $x - y$, which are proportional to a single exponential of the time, are called 'normal co-ordinates.' They lead on well to modifications at 45° to those of x and y .

When we are trying to interpret historical facts, we are not given that $\alpha = \beta$ and that $k = l$ exactly, although we may have reasons based on population, on industrialization, and on an assumption about common human nature, for believing that α/β and k/l are not far from unity. In these circumstances the normal co-ordinates would not be simply $x + y$ and $x - y$, but would be $l_1x + m_1y$ and $l_2x - m_2y$, where the l and m are positive weights. Now in an arms-race x is roughly proportional to y , and therefore $x + y$ is roughly proportional to $l_1x + m_1y$ for any not too different weights simply $x + y$. But there is no such automatic immunity from error in a replacement of $l_2x - m_2y$ by $x - y$; for they may even have different signs. This difficulty crops up in Sect. V below, where $x - y$ suits the First but not the Second World War.

All the varieties of motion that can be described by linear equations like (2/1) and (2/2), but without any restriction on the signs of α, β, k, l , have long been known (Lagrange, Weierstrass (31), 1875). They were considered, in the context of arms-races, by Richardson ((23), 1949, § 3/2/4); but none of them was found to resemble the phenomena of submissiveness, which must therefore require either variable coefficients, a higher order differential equation, an integral equation, or a non-linear description. That was my objection to a linear hypothesis with constant coefficients proposed by Mr. M. R. Horne as a representation of submissiveness ((9), 1951).

III. FIRST HYPOTHESIS ABOUT SUBMISSIVENESS

The statesmen of U.S.A., Britain, France, and U.S.S.R. still appear to pay no attention to mutual stimulation, for they take a pride in accumulating the type of weapons, such as bombing aircraft, or rockets, or submarines, which are most likely to provoke counter-measures. Perhaps they are relying on some unexpressed theory about submissiveness, thus: much force, much submission; therefore a show of force, some submission. That I call "thinking proportionately." The behaviour of living things is not always so simple. I remember an impertinent schoolboy who was excited and made more rebellious by various mild reproofs and punishments from a schoolmaster in the presence of a class of amused boys, but who was finally suppressed by a much more severe punishment. The behaviour of nations seems to involve a similar reversal.

In 1939 Richardson (21) modified (2/1) and (2/2) by non-linear terms intended to represent submissiveness, thus:

$$dx/dt = g - \alpha x + ky \{1 - \sigma(y - x)\}, \quad (3/1)$$

$$dy/dt = h - \beta y + lx \{1 - \rho(x - y)\}, \quad (3/2)$$

where σ and ρ were positive constants representing the tendency to submit.

Application of the First Hypothesis (a) To Very Unequal Nations.—Eqn. (3/1) and (3/2) were originally intended to represent a large x -nation overawing a much smaller y -nation. If we leave out g and h , as not of the essence of the matter, and if we neglect y in $x - y$, these equations become

$$dx/dt = -\alpha x + ky \{1 + \sigma x\}, \quad (3/3)$$

$$dy/dt = -\beta y + lx \{1 - \rho x\}. \quad (3/4)$$

It is dy/dt that will become negative. The smaller nation is suppressed.

For example in 1948 claims by Guatemala to British Honduras were suppressed by a show of armed force by Britain, without any fighting (K9153, K11511). I am not suggesting that the suppression of a smaller nation is necessarily either just or unjust; but am merely asserting that it is what usually happens in a world of power politics.

(b) *To Nations Equal as to their Constants.*—In 1951 I noticed that the terms representing submissiveness would have remarkable consequences even if the two opponents were equal in their constants. How this might happen was explained in a letter to *Nature* (1951, ix, 29) headed "Could an arms-race end without fighting?", in which arguments about increase or decrease were applied only in circumstances not far removed from the actual. Wishing, however, to see a general integral formula over the whole field, and not finding any, I applied to Dr. M. L. Cartwright, F.R.S., who has made a special study of non-linear differential equations ((4), 1949, (6), 1952). As an apology for the lack of general integral formulæ in the first two theories, I may quote her remark ((5), 1952, p. 88) that "what the pure mathematicians have done shows that the non-linear phenomena are genuinely complicated, and no easily applicable general theory can be expected." As the subject matter is economic or political, 3 per cent. accuracy coupled with a plainly intelligible meaning is to be preferred to infinite accuracy coupled with a complicated formula.

A theory of the submissiveness of equal groups must not be too deterministic: it must not say which of them will submit, for that is obviously a chancy affair. Such "stochastic processes" have been the subject of a symposium at the Royal Statistical Society (Moyal, J. E., Bartlett, M. S., and Kendall, D. G. (16), 1949).

By the time an arms-race has gone so far as to be alarming, former grievances and ambitions seem relatively unimportant; so that we may neglect g and h . On setting also $\alpha = \beta$, $k = l$, $\sigma = \rho$ eqn. (3/1) and (3/2) simplify to:

$$dx/dt = -\alpha x + ky + k\sigma(xy - y^2), \quad (3/5)$$

$$dy/dt = kx - \alpha y + k\sigma(xy - x^2). \quad (3/6)$$

From a study of the submission of Germany in the 1920s I estimated that:

$$\sigma > 0.34 \times 10^{-6} \text{ per person}, \quad (3/7)$$

(Richardson (23), 1949). As this information is meagre, let us change to new variables which involve σ in such a way that we can discuss its general effects without knowing its numerical value. Thus define U , V by

$$x = U/\sigma, \quad y = V/\sigma, \quad (3/8), (3/9)$$

so that U and V are pure numbers. Eqn. (3/5) and (3/6) then become

$$dU/dt = -\alpha U + kV + k(UV - V^2) \quad (3/10)$$

$$dV/dt = kU - \alpha V + k(UV - U^2). \quad (3/11)$$

$$t = \tau/k, \quad (3/12)$$

Further let

so that τ is a pure number. For thus

$$\frac{dU}{d\tau} = -\frac{\alpha}{k} U + V + UV - V^2, \quad (3/13)$$

$$\frac{dV}{d\tau} = U - \frac{\alpha}{k} V + UV - U^2. \quad (3/14)$$

Only one constant remains explicit, namely, α/k , and it is a pure number. Before 1877 arms-races did not occur, and statesmen believed in a stable balance of power, such as would occur if $\alpha/k \geq 1$. From a study of the arms-race which led to the First World War it was concluded that $k - \alpha =$

0.2, year⁻¹, whereas the disarmament after it gave $\alpha = 1.1_3$ years⁻¹ on the average (Richardson, 1949). Whence $\alpha/k = 0.8$. The disarmament after the Second World War gave $\alpha = 1.9$ years⁻¹. The present arms-race, with considerably more uncertainty, gives $k - \alpha = 2.2$ years⁻¹. These last two, taken together, give $\alpha/k = 0.46$ (Richardson (26), 1953). There is thus some evidence that α/k , has decreased, during the present century. The formulæ for any value of α/k are stated above; but in diagrams and in numerical examples the modern situation will be illustrated by

$$\alpha/k = \frac{1}{2}. \quad (3/15)$$

Arms-races between equal nations show most conspicuously in the sum of the warlike efforts of the two sides, whereas submission is essentially concerned with their difference. Let us therefore define S and D by

$$S = U + V, \quad D = V - U. \quad (3/16), (3/17)$$

By addition and subtraction of eqn. (3/13) and (3/14) it follows that

$$\frac{dS}{d\tau} = S \left(1 - \frac{\alpha}{k} \right) - D^2, \quad \frac{dD}{d\tau} = -D \left(1 + \frac{\alpha}{k} \right) + DS, \quad (3/18), (3/19)$$

or for a modern illustration

$$dS/d\tau = \frac{1}{2}S - D^2, \quad dD/d\tau = -\frac{3}{2}D + DS. \quad (3/20), (3/21)$$

We can still keep track of submissiveness by remembering that it comes in by way of the quadratic terms, and would cancel in the linear term.

These equations purport to represent a wide range of possible behaviour. In particular, suppose that initially the nations were considerably disarmed, as in 1948, or more so. The point $S = 0$, $D = 0$ is a point of balance. Sufficiently close to it the quadratic terms are negligible, so that

$$S = \text{const}_1 e^{\frac{1}{2}\tau}, \quad D = \text{const}_2 e^{-\frac{3}{2}\tau}. \quad (3/22), (3/23)$$

The exponential increase of S is the arms-race beginning inconspicuously. Meanwhile D tends to zero, so that the opposing sides tend to become equal in their warlike efforts. But (3/21) may be written

$$\frac{d \log |D|}{d\tau} = S - \frac{3}{2}. \quad (3/24)$$

As S grows, $\log |D|$ would at first become more negative; but there would come a time τ_1 , say, when $S = 3/2$, and immediately after that D would be unstable. That is how submissiveness would first begin to reveal its presence.

Too much determinism can be avoided by supposing that D had become near to zero; so that a chance disturbance could change its sign before or soon after it became unstable. After D^2 had grown sufficiently, $\frac{1}{2}S - D^2$ would become negative and so S would decrease. The decrease of S would be recognized as the end of the arms-race. The above argument was in effect that of my letter to *Nature* (25) of September 29th, 1951; and I still think it credible in the restricted part of the S, D plane to which it refers.

But now comes the difficulty. Dr. Cartwright, by sketching the integrals over a wider field, emphasized a theoretical point of balance from which the trajectories spiralled outwards. No corresponding historical facts have been found. There is the further difficulty that if the representative particle always comes out, how did it ever get in? There is a point of balance if simultaneously

$$0 = \frac{1}{2}S - D^2, \quad 0 = -\frac{3}{2}D + DS. \quad (3/25), (3/26)$$

The solution at $S = 0$, $D = 0$ was well known. It represented general disarmament, but was an unstable 'in-and-out' point. Other solutions are:

$$S = \frac{3}{2}, \quad D = \pm \frac{1}{2}\sqrt{3} = \pm 0.866. \quad (3/27), (3/28)$$

That is either (U, V) or $(V, U) = 1.183, 0.317$.

The general features of the theoretical field are shown in Fig. 1, which was obtained by the method used by B. van der Pol in solving his relaxation equation ((18), 1926). The lines, called 'isoclines,' along each of which dD/dS is constant, were drawn first. The isoclines are all conic sections, and each of them goes through both points of balance. For example, one of these conics is a pair of intersecting straight lines, each line passing through one point of balance. Points on each isocline were calculated and plotted, and a smooth curve was drawn through them. In contrast with that fairly definite construction, the drawing of the trajectories was in effect an integration, for it was a fitting together of pieces to form a continuous curve; and this process involved more uncertainty. However, van der Pol ((18), 1926) used it with good effect.

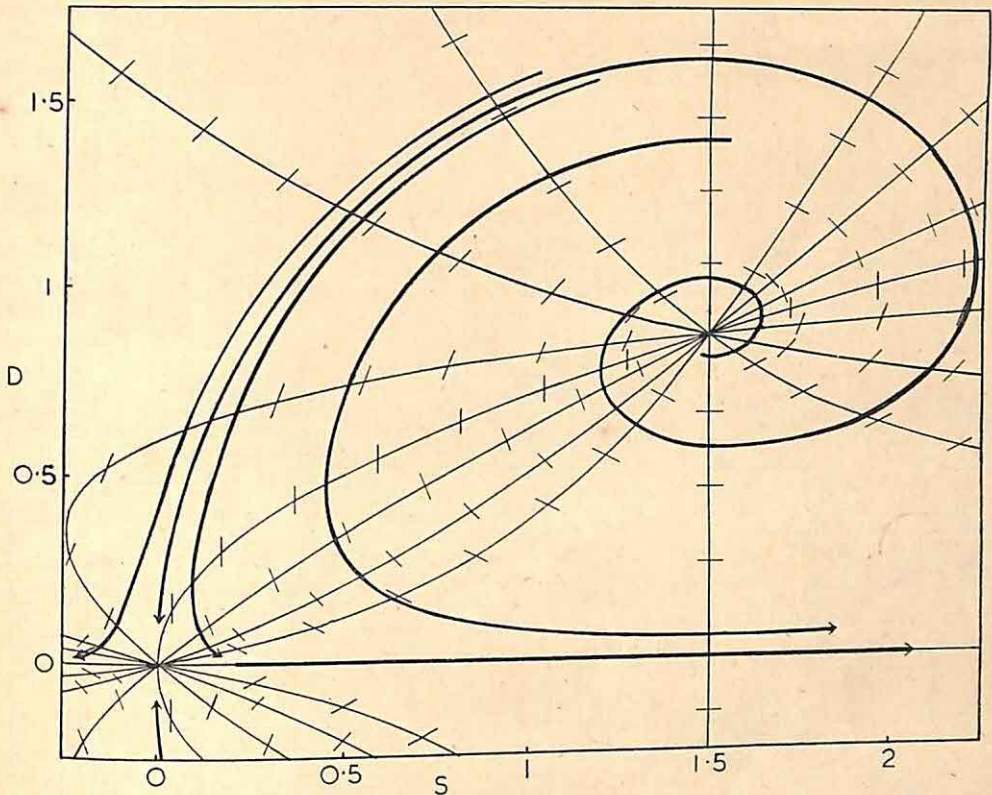


FIG. 1.—On the first hypothesis this shows the field of two opposing sides that are equal in their constants. S is proportional to the sum and D to the difference of their warfingersals. The thin curves are the isoclines along each of which dD/dS is constant. They are drawn for $dD/dS = 0, \frac{1}{2}, 1, \frac{3}{2}, \infty, -\frac{1}{2}, -1, -\frac{3}{2}$. The directions of motion are indicated by short segments crossing each isocline. The thick lines are trajectories drawn so as to fit the prescribed directions.

To investigate in more detail the neighbourhood of the point of balance at $S = \frac{3}{2}$, $D = \frac{\sqrt{3}}{2}$, define ε and ζ by

$$S = \frac{3}{2} + \varepsilon, \quad D = \frac{\sqrt{3}}{2} + \zeta, \quad (3/29), (3/30)$$

and substitute the expressions in the eqn. (3/20) and (3/21) obtaining

$$\frac{d\varepsilon}{d\tau} = \frac{1}{2} \varepsilon - \sqrt{3}\zeta - \zeta^2, \quad \frac{d\zeta}{d\tau} = \frac{\sqrt{3}}{2} \varepsilon + \varepsilon\zeta. \quad (3/31), (3/32)$$

The terms of zero degree have cancelled. On the circle $\varepsilon^2 + \zeta^2 = 1$ the terms of the second degree are of the same order of magnitude as those of the first degree. Much within that, say, inside the circle, $\varepsilon^2 + \zeta^2 = (1/10)^2$, the terms of the second degree become almost negligible. This circle of radius $1/10$, to the interior of which the following approximation is restricted, is quite a small part of Fig. 1. In the standard manner try the assumption that, where A, B are constants,

$$\varepsilon = Ae^{\lambda\tau}, \quad \zeta = Be^{\lambda\tau}, \quad (3/33), (3/34)$$

$$A\lambda = \frac{1}{2} A - \sqrt{3} B, \quad (3/35)$$

$$B\lambda = \frac{\sqrt{3}}{2} A. \quad (3/36)$$

then

and

These simultaneous equations for A and B are consistent if the determinant

$$\begin{vmatrix} \frac{1}{2} - \lambda & , & -\sqrt{3} \\ \frac{\sqrt{3}}{2} & , & -\lambda \end{vmatrix} = 0 : \quad (3/37)$$

that is, if

$$\lambda^2 - \frac{1}{2}\lambda + \frac{3}{2} = 0, \quad (3/38)$$

the roots of which are

$$\lambda_1, \lambda_2 = 0.25 \pm i 1.199. \quad (3/39)$$

Because these are complex with positive real part therefore the particle moves spirally outwards. The period of ε or ζ is $2\pi/1.199$ on the scale of τ . But $t = \tau/k$; and k has been found from the data of 1949, 1950, 1951 very doubtfully to be 4.1 years^{-1} . Therefore the period is :

$$\frac{2\pi}{1.199 \times 4.1} \text{ years} = 1.3 \text{ years}. \quad (3/40)$$

For a political double 'swing of the pendulum' that seems incredibly rapid. Maybe the warfingers for 1952 and 1953, when available, will show that k is less and the period greater. A method for making an accurate drawing of the trajectories in the neighbourhood of any point of balance is worked out in the author's *Arms and Insecurity* ((23), 1949, § 3/2/4). When the roots λ are complex, the trajectory can be obtained from an auxiliary equiangular spiral by stretching or squeezing it at right angles to a characteristic direction E . Application of those general formulæ to the present case shows that the radius vector of the equiangular spiral revolves counter-clockwise and that its length increases in the ratio 3.707 per turn, or in the ratio 1.388 per quarter turn. The characteristic direction is at $E = 105^\circ$ from the ε -axis. The trajectory is obtained by stretching the equiangular spiral in the ratio 1.5005.

Dr. Chike Obi (17), of University College, Ibadan, Nigeria, who was formerly a pupil of Dr. Cartwright, has made an extensive study, from the point of view of pure mathematics, of the equation

$$\ddot{x} + (\alpha + \beta x) \dot{x} + x + \gamma x^2 = 0, \quad (3/41)$$

where α, β, γ are constants not too far removed from zero. It is possible by a chain of transformations to connect the present hypothesis exactly to Obi's form. For this purpose we first introduce a new variable $\omega = D^2$, then shift the origin to the vortex point, and finally eliminate that deviation which is not in the S direction.

It is sometimes asserted that 'history repeats itself.' If so, we should see a closed orbit. Actually there are only spirals in Fig. 1.

The origin is what I have elsewhere (1949) named an 'in-and-out' point. Lefschetz ((12), 1948) names it a 'col.' Cartwright calls it a 'saddle point' or 'col.' To the right of it the drift is towards rearmament; to the left of it the drift is towards negative armament, which I formerly suggested ((21), 1939, p. 7) was another name for co-operation. The practical difficulty was: how could the nations ever get across to the co-operative side, seeing that the flow diagram does not permit a direct way into the co-operative region. These are strange ideas. They all purport to refer to times when there was no fighting.

There is a boundary-spiral which runs right into the 'in-and-out' point. Any slightly lesser spiral passes first through the arms-race. Any slightly greater spiral goes beyond the 'in-and-out' point into the co-operative region.

This vortex, which certainly follows from hypothesis I, is very difficult to accept as a description of what nations in fact did. Herein lies a good example of the usefulness of mathematics in social science. The consequences of the hypothesis were not all immediately obvious. In the familiar region they seemed plausible. It was only by deducing consequences in an unfamiliar region that incredible entailments were noticed.

Let us therefore pass on to a different hypothesis.

IV. SECOND HYPOTHESIS ABOUT SUBMISSIVENESS

As in the first hypothesis, the defence-coefficient must become negative when the threat becomes overwhelming; but this change of sign will now be contrived differently. One of the reasons given ((21) 1939, p. 23) for the theory which has just been abandoned was that "the

overwhelmingness of the threat depends not on y absolutely, but on y relative to x ." The reconsideration is that it is not relative to the x present at that time, but to the x that could be mobilized. This ' x that could be mobilized' is almost a constant for any one nation, but varies very much from one nation to another.

Thus let the new hypothesis be

$$dx/dt = g - \alpha x + ky \{1 - y/m\}, \quad (4/1)$$

$$dy/dt = h - \beta y + lx \{1 - x/n\}, \quad (4/2)$$

where m and n are constants for those nations, and are such that m is proportional to the total warfiness of the x -nation could mobilize, and n is proportional to the corresponding quantity for the y -nation. Therefore

$$x \leq m, \quad y \leq n. \quad (4/3), (4/4)$$

As before, g and h will be neglected, because they are not essential to the phenomena of submissiveness.

Consequences of the Second Hypothesis (a) For Very Unequal Nations.—Suppose that $m \gg n$, then n/m , and *a fortiori* y/m , is near zero: so that $1 - y/m$ is almost unity. There is, on the contrary, no reason why x/n should not much exceed unity. In these circumstances we have approximately from (4/1) and (4/2)

$$dx/dt = -\alpha x + ky, \quad (4/5)$$

$$dy/dt = -\beta y - lx^2/n. \quad (4/6)$$

Therefore y is decreasing. The smaller nation is suppressed. For this case of very unequal nations there is nothing much to choose between the two hypotheses. Either of them describes what usually happens in a world of power politics.

(b) *For Nations Equal in their Constants.*—From (4/1) and (4/2) we have

$$dx/dt = -\alpha x + ky(1 - y/m), \quad (4/7)$$

$$dy/dt = -\alpha y + kx(1 - x/m). \quad (4/8)$$

The constant m can be made to disappear by a change of variables. Thus define u, v by

$$x = um, \quad y = vm; \quad (4/9), (4/10)$$

so that

$$u \leq 1, \quad v \leq 1. \quad (4/11), (4/12)$$

For then

$$du/dt = -\alpha u + kv - kv^2, \quad (4/13)$$

$$dv/dt = -\alpha v + ku - ku^2. \quad (4/14)$$

The two remaining constants, α and k , can be reduced to one, in the following way:

As in the former theory, define τ by

$$t = \tau/k; \quad (4/15)$$

so that

$$\frac{du}{d\tau} = -\frac{\alpha}{k}u + v - v^2, \quad (4/16)$$

$$\frac{dv}{d\tau} = -\frac{\alpha}{k}v + u - u^2. \quad (4/17)$$

Finally, define S and D by

$$S = u + v, \quad D = v - u, \quad (4/18), (4/19)$$

obtaining

$$\frac{dS}{d\tau} = \left(1 - \frac{\alpha}{k}\right)S - \frac{1}{2}(S^2 + D^2), \quad (4/20)$$

$$\frac{dD}{d\tau} = -\left(1 + \frac{\alpha}{k}\right)D + SD. \quad (4/21)$$

Or, with $\alpha/k = \frac{1}{2}$ as in the former theory, to represent modern conditions

$$dS/d\tau = \frac{1}{2}S - \frac{1}{2}(S^2 + D^2), \quad (4/22)$$

$$dD/d\tau = -\frac{3}{2}D + SD. \quad (4/23)$$

The equation for $dD/d\tau$ is exactly the same as on the former hypothesis; but that for $dS/d\tau$ has $\frac{1}{2}(S^2 + D^2)$ in place of simply D^2 . The points of balance where $dS/d\tau$ and $dD/d\tau$ vanish simultaneously are given by

$$0 = S - S^2 - D^2, \quad 0 = D \left(S - \frac{3}{2}\right). \quad (4/24), (4/25)$$

If from (4/25) we accept $D = 0$ and substitute in (4/24) we get $0 = S - S^2$ which has solutions $S = 0$ or $S = 1$. Alternatively, if $S = 3/2$, the second condition is satisfied and the first becomes $D^2 = -3/4$, so that D is imaginary. There are therefore only two points of balance, namely,

$$S = 0, D = 0 \text{ and } S = 1, D = 0. \quad (4/26), (4/27)$$

We seem well rid of the puzzlingly unsymmetrical point of balance of the former theory. The point $S = 0, D = 0$ is of course of the same character as before, namely, an in-and-out point.

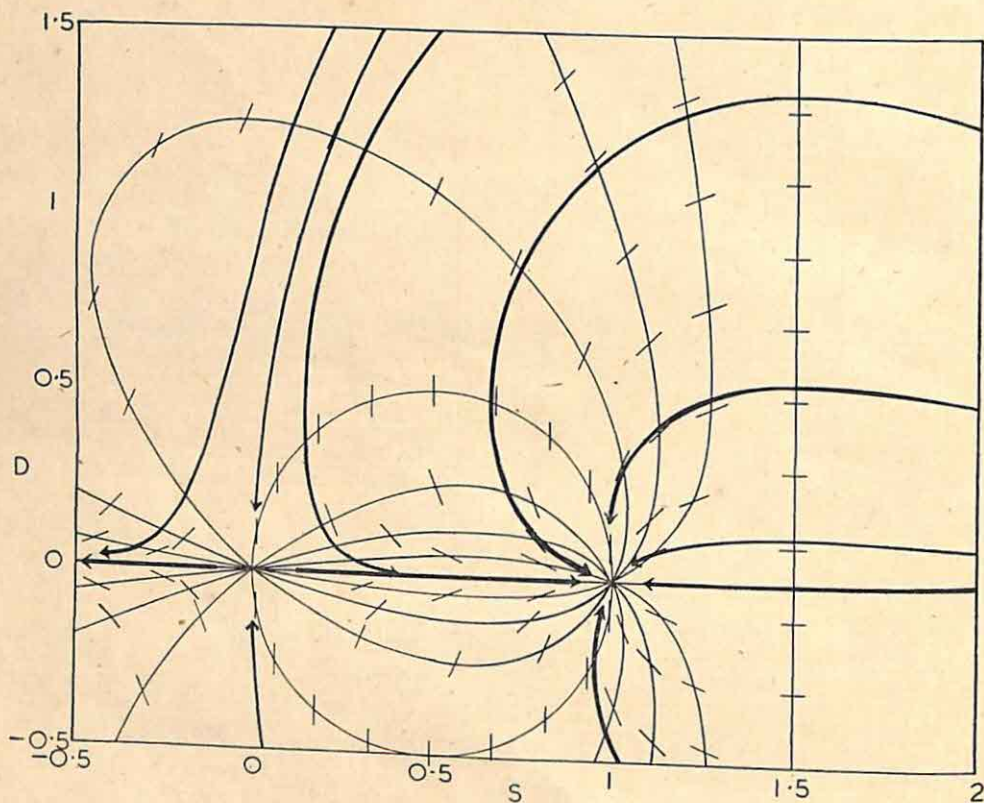


FIG. 2.—The field of two opposing sides on the second hypothesis. As in Fig. 1, S is proportional to the sum and D to the difference of their warfinpersals. The sides are equal as to their constants. The thin curves are the isoclines, along each of which dD/dS is constant. The direction of motion is indicated by a short segment. The thick curves are trajectories, along any one of which the point which represents the two warfinpersals could move.

The general field is shown in Fig. 2 by way of isoclines and trajectories. The isocline

$$\frac{dD}{dS} = j, \quad (4/28)$$

where j is any constant, works out to

$$S - S^2 - D^2 = \frac{D}{j} (2S - 3); \quad (4/29)$$

so that changing the sign of j has the same effect as changing the sign of D .

In the region of the arms-race, where D tends to zero while S increases, we may neglect D in (4/22) which then simplifies to

$$\frac{dS}{d\tau} = \frac{1}{2} S - \frac{1}{2} S^2, \quad (4/30)$$

in which the variables are separable. Its integral can be verified to be

$$\frac{1}{S} = 1 + Ce^{-\tau/2}, \quad (4/31)$$

where C is an arbitrary constant. So as $\tau \rightarrow \infty$, $S \rightarrow 1$. Let us investigate the neighbourhood of the point $S = 1$, $D = 0$ on the second hypothesis. Define ε here differently from the ε in (3/29), and thus

$$S = 1 + \varepsilon. \quad (4/32)$$

Insert this in (4/22), (4/23). Then

$$d\varepsilon/d\tau = -\frac{1}{2}\varepsilon - \frac{1}{2}(\varepsilon^2 + D^2), \quad (4/33)$$

$$dD/d\tau = -\frac{1}{2}D + \varepsilon D. \quad (4/34)$$

If we neglect the quadratic terms in ε and D , the solutions are

$$\varepsilon = Ae^{-\tau/2}, \quad D = Be^{-\tau/2}, \quad (4/35), (4/36)$$

where A, B are arbitrary constants. This point is therefore one of thoroughly stable balance.

The theoretical state of affairs at which we have arrived as a deduction from the second hypothesis might please some people: large armed forces, stable balance, and therefore no fighting. One of the editors of the *Cambridge Modern History* (3), Stanley Leathes, wrote in his preface to Vol. XII: "On the whole, the existence of this tremendous military equipment makes for peace. The consequences of war would be felt in every household; and statesmen, as well as nations, shrink from the thought of a conflict so immense." This was published in 1910. Unfortunately for Leathes' theory, the First World War broke out four years later. Yet this theory persists. Winston Churchill, speaking at the Pilgrims' dinner on October 14th, 1952, said that "a third world war is unlikely because both sides know that it would begin with horrors never dreamt of before."

So my second hypothesis about submissiveness appears to agree with Leathes and with Churchill, but to have the outbreak of the First World War as evidence against it. Personally I have no confidence in it, regarding it as wishful thinking.

V. THIRD HYPOTHESIS (FOR SIDES EQUAL IN THEIR CONSTANTS)

When formulating the two former hypotheses, I had hoped to have some insight into what the motives might be. Now I ignore motives, and try to make a mathematical description of the courses in time of the opposing warfingers before, during, and after the First World War. The facts must first be reviewed.

(i) From A.D. 1907 to 1914, the two sides being the Triple Entente and the Triple Alliance, the representative point was moving nearly in a straight line with a speed proportional to its distance from a "point of balance" at $x = £135$ million, $y = £73$ million. The graph is given by Richardson ((26), 1953). This is called an arms-race. The sterling has been converted to warfingers by division by £97.3 per year; that being a British wage the conversion is very rough for the other nations. Before these numbers were plotted in Fig. 3, x and y were interchanged.

(ii) There followed the First World War. Italy changed sides. The warfinger of Britain attained a maximum of 12.2 million equivalent persons in 1917 (Richardson (26), 1953). I have not succeeded in directly ascertaining the warfingers of the other belligerents, but a rough indirect estimate can be made as follows. Lloyd George ((13), *War Memoirs*, 1938, p. 1575) states the composition of the British Army in March, 1918 (excluding Colonial and Indian troops). From this statement I compute that there were 2.96 million British combatants. Now $12.2/2.96 = 4.13$. It is understandable that the warfinger might considerably exceed the combatant strength, because the warfinger accounts for the non-combatants and the munition makers. Lloyd George (p. 1,577) further states that at the end of 1917 the total combatant strength of the Allies, without including any Russians or Roumanians, was 5.4 million against 5.2 million for the Central Powers. My rough estimate is made by assuming that the British factor, 4.13, applied also to the other nations concerned, so that the maximum warfingers were: Allies 22.3×10^6 , Central Powers 21.5×10^6 , total 43.8×10^6 equivalent persons.

Because Germany ultimately submitted, it is relevant to enquire when the first hint of submissiveness appeared in the German state as a whole. The question is not about individuals or

groups. In another place ((22, 1948) I have given a theory of war-moods in which the people in German groups are represented by five variables; but here only one variable is available. On December 12th, 1916, the German Chancellor suggested a peace conference. The opposing Powers, after consultation, said they "refuse to consider a proposal which is empty and insincere." (Lloyd George, p. 661.) There were various other tentative and secret peace feelers in 1917 and 1918, but mostly Austrian. According to a Reichstag Committee, quoted by Lloyd George (p. 1932), "Up to 15th July, 1918, the Supreme Army Command rejected the view that victory was no longer possible of attainment by force of arms, and gave no support to peace negotiations upon the basis of a military stalemate . . ." During the next three weeks the fresh Americans came to the aid of the tired French and British. A battle on August 8th, 1918, in which the Germans had no reserves left, changed Ludendorff's opinion, and on August 14th the Kaiser, presiding over a Crown Council at Spa, admitted that Germany would have to find a suitable moment in which to come to an understanding with the enemy.

(iii) After the First World War, the sides were not quite the same. There was the Russian revolution and its consequences. U.S.A. withdrew into isolation. A new diagram was drawn showing the warfingers of Germany plotted against that of its entourage, consisting of France, Britain, Poland, Czechoslovakia, and Russia. This appears in a microfilm (Richardson (23), 1949) and is modified here into part of Fig. 3. It shows motion, along a nearly straight track and slowing down as a point of balance is approached. There was a delay of eleven years in the neighbourhood of this point, and then a departure, with increasing speed along another almost straight track. This was the second great arms-race beginning.

(iv) It led to the Second World War during which the warfingers approached the state of total mobilization, but were difficult to ascertain. The sides were rather different from those of the first war, Japan and Italy in particular having changed over previously, and Italy changing back again. In October, 1939, Hitler made a peace offer: but Britain and France did not regard it as at all submissive. After that the demand for the unconditional surrender of Germany delayed any sign of submissiveness until defeat.

(v) The Occupation Statute for Germany, an expression of submission, has now been in operation for seven years. Disarmament after the Second World War was barely over when enmity between U.S.A. and the Soviet Union began to dominate the scene.

Fitting Formulae to the Facts.—It seems permissible to ignore the changes in the membership of the sides, so that x and y may be regarded as having the same meaning throughout. This has already been assumed in Fig. 3.

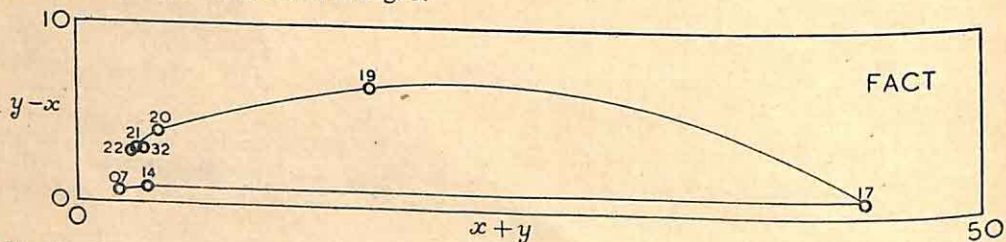


FIG. 3.—Pro- and anti-German warfingers in Europe, 1907 to 1932, simplified by ignoring other changes in the sides; and in particular by ignoring the brief, but balance-altering, intervention of U.S.A. The co-ordinates are the sum and difference. The numbers set beside the plotted points are the dates.

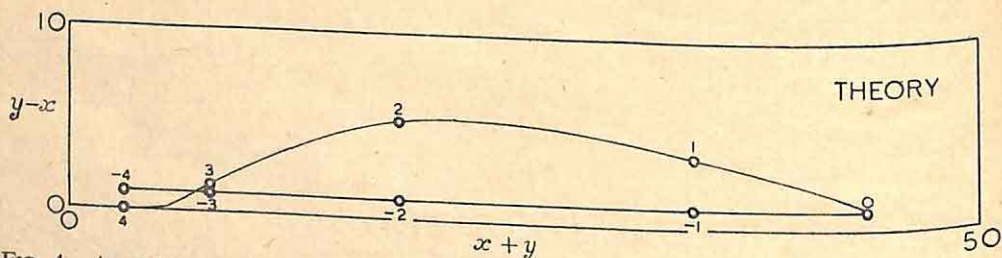


FIG. 4.—A mathematical representation, plotted from formulae (6/17), (6/19), and (6/20) of the warfingers before, during, and after the First World War. For comparison with Fig. 3. Numbers set beside the plotted points are values of t . A unit of one year makes Fig. 4 agree roughly with Fig. 3 as to time.

As the mathematical description is to be of that which usually happened, we are led to beg the question "Could an arms-race end without fighting?" by assuming, on the meagre evidence of two out of three trials, that it usually does not. This assumption is in strong contrast with the first and second hypotheses, which leave that important question open at the outset.

Remembering that an arms-race has to do mainly with the sum $x + y$ and submission mainly with the difference $y - x$ of the warfingersals, let ξ and η be defined thus:

$$\xi = x + y, \quad \eta = y - x, \quad (6/1), (6/2)$$

so that conversely

$$x = \frac{1}{2}(\xi - \eta), \quad y = \frac{1}{2}(\xi + \eta). \quad (6/3), (6/4)$$

Accordingly ξ and η are respectively constant multiples of the S and D of former theories. We can now describe the changes in ξ without mentioning η , thus:

- (i) After wars there have been disarmaments. Therefore $d\xi/dt$ must be capable of either sign.
- (ii) For small ξ it is observed that $d\xi/dt$ varied as $\pm(\xi - \xi_a)$ where ξ_a is a constant, called a point of balance.
- (iii) There was an upper bound to ξ when both the opposing sides were fully mobilized for war. Let γ denote this upper bound.

The above three requirements are satisfied by the assumption that

$$\left(\frac{d\xi}{dt}\right)^2 = a^2(\xi - \xi_a)^2(\gamma - \xi), \quad (6/5)$$

where a^2 is some constant. A simpler expression could not be found; η does not appear; ξ cannot exceed γ , for if it did then $d\xi/dt$ would become imaginary; when ξ is near zero then $d\xi/dt \approx \pm a(\xi - \xi_a)\sqrt{\gamma}$.

The variables are separable, so that (6/5) can be integrated. The result, which can be verified by differentiation, is

$$\xi - \xi_a = (\gamma - \xi_a) \operatorname{sech}^2 Bt, \quad (6/6)$$

$$B = \frac{a\sqrt{(\gamma - \xi_a)}}{2}. \quad (6/7)$$

where

At $t = 0$, $\xi = \gamma$; that is to say, the constant of integration has been chosen so that $t = 0$ at the maximum of ξ , which occurred during the war. As $t \rightarrow -\infty$, $\xi \rightarrow \xi_a$: there is not, on this theory, any sudden or definite beginning of the increase of warfingersal. This vaguely agrees with the author's general impression of a collection of deadly quarrels that ended since A.D. 1820 (Richardson (24), 1950). The question "Who began it?" seldom had any non-controversial answer. He abandoned any attempt in that collection to name the aggressors.

To determine B from observation, let t' be the time at which ξ was halfway between γ and ξ_a . It follows from (6/6) that

$$B = \frac{0.8814}{t'}. \quad (6/8)$$

The theoretical diagram, Fig. 4, has been drawn with

$$B = 0.5 \text{ year}^{-1}. \quad (6/9)$$

Next as to η , the facts represented in Fig. 3 suggest that η did not alter much while ξ was increasing, but that, after the maximum of ξ , changes in η were conspicuous. These were of two kinds: at first, while ξ was large, $|\eta|$ increased; later when ξ was near ξ_a there was a settling of η to about the same value ξ_a , so that finally

$$x = \frac{1}{2}(\xi - \eta) = 0, \quad (6/10)$$

$$y = \frac{1}{2}(\xi + \eta) = \xi_a. \quad (6/11)$$

This expresses the final submission of the x -side. It would have been a mistake to have made η tend to zero, unless ξ also did. To express both kinds of change in η we need something like a disjunction or "either-or" formula. One can be concocted on the model of $\xi + 1/\xi$ which behaves like ξ when ξ is large, but like $1/\xi$ when ξ is near zero. When ξ was large, η tended from zero without changing sign, as if $d\eta/dt$ were proportional to η . When ξ was small, η tended to ξ_a , as if $d\eta/dt$ were proportional to $\xi_a - \eta$. These diverse tendencies could be combined by weighting them respectively by ξ and $1/\xi$. That, however, would make the two terms of different

dimensions, for ξ is a number of equivalent persons. So also is γ . It is found to be satisfactory to use the dimensionless weights $2\xi/\gamma$ and $\gamma/2\xi$. Thus one arrives at the assumption that

$$\frac{d\eta}{dt} = \sigma_3 \left\{ \eta \left(\frac{2\xi}{\gamma} \right) + (\xi_b - \eta) \left(\frac{\gamma}{2\xi} \right) \right\}, \quad (6/12)$$

in which σ_3 is some constant measure of submissiveness, and is the reciprocal of a time.

A most important fact about the submissiveness of Germany as a whole was that it came on suddenly. Sudden commencements have been much studied mathematically in connexion with seismographs, and with the switching on of electric circuits. Heaviside introduced a function $H(t)$ defined to be zero for $t < 0$, but unity for $t > 0$ (H. and B. S. Jeffreys (30), 1946, p. 219). If the onset is required to be not quite so sudden, four approximations to $H(t)$ are available (Van der Pol and Bremmer (19), 1950, pp. 56-7). I shall write

$$\frac{d\eta}{dt} = H(t) \sigma_3 \left\{ \eta \frac{2\xi}{\gamma} + (\xi_b - \eta) \frac{\gamma}{2\xi} \right\}. \quad (6/13)$$

So that $d\eta/dt = 0$ for $t < 0$.

The inevitable non-linearity has been introduced, but in a manner quite different from that of hypotheses I and II.

As a discontinuity at $t = 0$ in $d\eta/dt$ has been admitted, a discontinuity also in $d^2\xi/dt^2$ does not seem too fanciful. It allows us to joint together at $t = 0$ two formulæ for ξ , both like (6/6), but the later one having b for a and C for B , so that the final point of balance ξ_b need not agree with the initial ξ_a . Observed constants, in million equivalent persons, were about

$$\gamma = 44, \quad \xi_b = 3. \quad (6/14), (6/15)$$

The integration of (6/13) is, however, much simplified if we take $\xi_b = 0$, for then η is a factor of the second member, so that for $t > 0$

$$\frac{d \log |\eta|}{dt} = \sigma_3 \left\{ \frac{2\xi}{\gamma} - \frac{\gamma}{2\xi} \right\}. \quad (6/16)$$

This simplified problem is solved as a preliminary. Substitute, in view of (6/6) and (6/9),

$$\xi = \gamma \operatorname{sech}^2 \frac{t}{2}, \quad (6/17)$$

obtaining from (6/16)

$$\frac{d \log |\eta|}{dt} = \sigma_3 \left\{ 2 \operatorname{sech}^2 \frac{t}{2} - \frac{1}{2} \cosh^2 \frac{t}{2} \right\}. \quad (6/18)$$

The integral is

$$\frac{\eta}{\eta_0} = \operatorname{Exp} \left[\sigma_3 \left\{ 4 \tanh \frac{t}{2} - \frac{1}{4} \sinh t - \frac{1}{4} t \right\} \right], \quad (6/19)$$

which can be verified by differentiation. The measure σ_3 of submissiveness was, of course, unknown ; but Fig. 4 was plotted with

$$\sigma_3 = 1 \quad (6/20)$$

and comparison of Fig. 4 with Fig. 3 shows that $\sigma_3 = 1$ is of the right order of magnitude. The complete eqn. (6/13), with ξ_b not zero, has for an integrating factor the expression (6/19) (Forsyth (7), 1929, p. 20) and is thus reducible to quadratures. But these look difficult and are perhaps not worth pursuing at this stage.

VI. CONCLUSION

The memory of recent overwhelming military defeat, say, within five years, certainly leaves a tendency to submit. A large state has been known to overawe a very much smaller state, without any fighting. Apart from those two types, I have looked for submissive effects, but have not noticed any. Two hypotheses as to how submissive effects might arise between rested and approximately equal states, called A and B, are considered. Mathematical hypotheses like these are too complicated to be expressed, both briefly and accurately, in words. However, one can say that the description of an arms-race, as caused by mutual stimulation, is modified by supposing either (i) that the overwhelming nature of the threat

against A is proportional to the excess of B's armed force over A's at the same time, or (ii) that it is not A's armed force at the same time that matters, but what A could mobilize. Not all the consequences of these hypotheses are immediately obvious. In historically known circumstances the hypotheses seemed plausible. It was only by deducing their consequences in unusual circumstances that incredible entailments were noticed. This is a good example of the usefulness of mathematics in social science.

Most governments seem to expect that a show of armed force, to a rested and equal group of nations, will have an effect of the same sign as an overwhelming military defeat. This expectation appears to be a mistake: the actual effects go opposite ways.

The Quakers, Gandhi, and others have tried to persuade everybody to be as non-violent as the exceptional Christian saints. Most people feel the suggestion to be too utterly contrary to their defensive impulses. Could not a middle way be found, with the motto "defence without provocation." This motto might well be applied to words as well as deeds. The existence of the various arms of nation A presumably have different psychological effects in peacetime on nation B; for example, one may guess that B is much more provoked to retaliate by the existence of A's long-range bombers than by A's land-based fighters. Jonathan Griffin (8) discussed 'defence without menace' in 1936, and now again the problem of the classification of armaments into provocative and non-provocative requires much more attention than it receives.

Submission, except in the sense of defeat, is not a spectacular subject for newspaper headlines; it is not like fighting for freedom. Yet we need to remember that in any organization, whether it be a social club, a scientific society, a business, a national parliament, or the United Nations, a suitable amount and distribution of submissiveness is essential for smooth working. If each member struggled all the time for his own advantage there would be chaos.

The present study has perhaps over-emphasized the bitter submission of the defeated, threatened, suppressed, or snubbed, and has neglected the joyous devotion to an admired leader. To correct the balance let me instance a classical story illustrating the latter tendency, that of Naomi and Ruth. It is described as an affair of tribal loyalties.

In the United Nations, activities of the follow-my-leader type are similarly noticeable on both sides of the major schism.

REFERENCES

1. Allport, G. W. (1945). *Personality*. New York: Henry Holt.
2. Bateson, G. (1936). *Naven*. Cambridge: University Press.
3. *Cambridge Modern History*, XII (1910). Cambridge: University Press.
4. Cartwright, M. L. (1949). Report on Non-Linear Vibrations. *The Advancement of Science*, 6.
5. Cartwright, M. L. (1952). *Non-Linear Vibrations*. *Math. Gazette*, 36, 81-8.
6. Cartwright, M. L. (no date). *Forced oscillations in non-linear systems*. (Lectures at Professor Lefschetz's seminar.)
7. Forsyth, A. R. (1929). *A Treatise on Differential Equations*. London: Macmillan.
8. Griffin, J. (1936). *Alternative to Rearmament*. London: Macmillan.
9. Horne, M. R. (1951). Letter in *Nature*, November 24th.
10. *Keesing's Contemporary Archives*. Bristol: Keesing's Publications: cited as 'K.'
11. Landau, H. G. (1951). *Bull. Math. Biophysics*, 13, 1-19.
12. Lefschetz, S. (1948). *Lectures on Differential Equations*. Princeton: University Press.
13. Lloyd George, D. (1938). *War Memoirs*. London: Odhams.
14. Lorenz, K. Z. (1952). *King Solomon's Ring*. London: Methuen.
15. McDougall, W. (1917). *Social Psychology*. London: Methuen.
16. Moyal, J. E., Bartlett, M. S., and Kendall, D. G. (1949). 'Symposium on Stochastic Processes.' *J. Roy. Stat. Soc., B*, 11, 150-282.
17. Obi, C. (1950). (Unpublished.) Thesis for Ph.D.: Cambridge University.
18. Pol, B. van der (1926). *Phil. Mag.*, II, 978-92.

19. Pol, B. van der, and Bremmer, H. (1950). *Operational Calculus based on the Two-sided Laplace Integral*. Cambridge: University Press.
20. Richardson, L. F. (1935). *Nature*, **135**, 830.
21. Richardson, L. F. (1939). 'Generalized Foreign Politics.' *Brit. J. Psychol. Mon. Sup.*, No. 23.
22. Richardson, L. F. (1948). *Psychometrika*, **13**, 147-74 and 197-232.
23. Richardson, L. F. (1949). *Arms and Insecurity* (microfilm, published by the author).
24. Richardson, L. F. (1950). *Statistics of Deadly Quarrels* (microfilm, published by the author).
25. Richardson, L. F. (1951). *Nature*, **168**, 567.
26. Richardson, L. F. (1953). *The Indian Journal of Statistics*, **12**, 205-28.
27. Richardson, L. F., and Stommel, H. (1948). *Journ. Meteor.*, **5**, 238-40.
28. Russell, Bertrand (1914). *War the Offspring of Fear*. London: Union of Democratic Control.
29. Volterra, V. (1930). *Theory of Functionals*. London: Blackie.
30. Jeffreys, H. and B. S. (1946). *Methods of Mathematical Physics*. Cambridge: University Press.
31. Weierstrass, K. (1875). *K. Akad. der Wiss.* (October 28th). Also *Werke*, II, 75-6.

A PRELIMINARY STUDY OF THE THEMATIC APPERCEPTION TEST

By AMYA SEN

University College, London

I. *History of Method.* II. *Subjects.* III. *Test Material and Mode of Scoring.*
IV. *Reliability and Validity of Assessments.* V. *Summary and Conclusions.*

I. HISTORY OF METHOD

In what is known as the 'thematic apperception test' a series of somewhat vague and problematic situations are presented to the subject, and he is asked to make up a brief story explaining the situation, and indicating the feelings and thoughts of the principal characters. In the more usual form the material consists of standard pictures; in other forms the situations are described verbally. The psychologist then attempts to infer the personal characteristics of the individual tested from the nature of the 'themes' appearing in his stories.

The main object of the investigation described in the following pages was (i) to examine the possibility of scoring the pictorial form of the test in accordance with a more adequate theory of personality, and in particular (ii) to estimate the validity of assessments so obtained for certain basic traits.

In order to decide which traits are likely to lend themselves to assessment by such a test, it is essential to begin by asking what kind of mental activities such tests tend, in general, to provoke. This preliminary study has been strangely neglected by nearly all who have devised or used such procedures.¹

The use of pictorial material for tests of character and personality has a far longer history than is commonly supposed. As long ago as 1886, Dr. Sophie Bryant,² the first headmistress of a famous London High School and one of Francis Galton's most active disciples, carried out what must have been the earliest set of experiments in testing character. As stimuli she used both pictures and words, and interpreted the answers chiefly in terms of the crude associationist theory which was then current (1). Binet, in his studies of individual differences, got children to describe both pictures and objects, and distinguished four main types—the observer, the describer, the erudite, and the emotional ((3) and other papers); later he included pictures in his scale of intelligence tests (4). Most of the older manuals of experimental psychology include experiments with ambiguous pictures, diagrams, or ink blots, designed to illustrate the principles of apperception;³ and nearly all emphasize the importance of personal and emotional factors,⁴ as well as tendencies not unlike those formerly described as 'primary and secondary laws of association.'⁵

¹ For much of what follows I am indebted to roneo'd Laboratory Notes by Professor Burt (15).

² Curiously enough, Vernon, in his survey of British work on personality tests, refers to none of the various papers published by Sophie Bryant. There are several references to her work in Whipple's account of 'tests of description and report' ((22), pp. 290, 292, 313, 319).

³ E.g., Witmer, L., *Analytical Psychology: A Practical Manual* (1902), experiments 1-8; Seashore, C. E., *Elementary Experiments in Psychology* (1908), ch. xii.

⁴ "Apperception," says Stout (and this is the special interpretation developed by British in contrast to continental writers), "is essentially a *conative* process; . . . it is the process by which a conation fulfils itself" (*Analytic Psychology*, II, 1896, pp. 114, 121); it is instructive to note how Stout's discussion of the 'conflict of apperceptive systems' (pp. 145-62) anticipates the Freudian doctrine of conflict, repression, and the influence of unconscious tendencies, and indicates their importance for the organization of personality.

⁵ Cf. Stout, *op. cit.*, pp. 132f., on the 'conditions determining the strength of apperceptive systems.'

It was a growing realization of the defects of the older association theories that led British writers to develop the idea of a more general integrative principle: in doing so, their criticisms of associationism and their emphasis on organization rather than mere linkage largely anticipated the views of the Gestalt psychologists, particularly in the field of personality. The introduction of the term and idea of 'apperception' into modern psychology is due chiefly to the Herbartian school, particularly in this country to Stout (2) and Adams (5). It may perhaps best be defined as "the process whereby the content of a particular sensory presentation, at first relatively indeterminate and meaningless, seems to become clearer and more determinate, and so acquires for the perceiver some kind of apparent meaning, in virtue of a fusion or synthesis with contents already systematically organized in the perceiver's mind." All Herbart's followers insist that a man's mode of apperceiving an object depends largely on his own individual 'apperception-systems,' and hence may incidentally provide a clue to his intelligence, interests, and character. Adams relates the legend of the Indian princess, who embroidered on her banner representations of herself, a coin of the realm, and her father's imperial crown, all superposed to make a single picture-puzzle; each of her suitors was required to say what the device portrayed: the lover who saw, not the money or the crown, but the portrait won the contest. Steinthal, one of Herbart's best-known exponents, quotes a similar story: "we place a man," he concludes, "largely by the point of view from which he regards things, in short, by his mode of apperceiving."¹

Binet, Stern, and most of the earlier investigators who used pictorial tests were interested chiefly in the examinee's ability to observe what is before him and to organize what he observes into an intelligent report. The first to exploit the principle of apperception for testing the dynamic aspects of personality appears to have been Burt. In describing his tests of 'emotional activities' he writes: "Ambiguous pictures, that is, pictures which can be *apperceived* or attended to in various incompatible ways, are exhibited to the different individuals; . . . and the particular mode of apperception shows consistent and significant differences." Some of the most conspicuous peculiarities were those revealing the distinctive interests of the two sexes (7). He observes that hitherto most writers had accepted Whipple's definition of a 'mental test' as denoting 'the experimental determination of some phase of mental capacity' ((6), p. 1), thus restricting it to cognitive processes; the attempt to measure the affective and conative processes, or, as he sometimes terms them, the 'dynamic or orectic aspects' of personality, has been "almost entirely neglected." He reports the results of various experimental procedures, and concludes that "the best tests of emotional disposition are to be obtained by the method of *impression* rather than of *expression*," i.e., by their influence on such processes as attention, association, preference, and apperception ((7), p. 381; cf. (9)).

Since the methods of scoring and interpretation adopted in the present investigation with adults were based on those worked out by Burt in his work with children, it may be helpful to describe quite briefly his general procedure and how the results were classified and marked.

In his re-standardization of the Binet intelligence-tests he notes how Binet's picture of a mysterious figure standing on a couch and looking out of a window elicited striking individual differences, quite apart from variations in observation or intelligence: "some children describe the man more or less objectively as he appears in the picture; others subjectively identify themselves with the figure, and describe what he might be supposed to see—basing their descriptions largely on their own wishes, experiences, or fears." He accordingly retained it in his series of 'problem pictures.' The others used in these earlier researches were: Watts' *Hope*, Bramley's *A Hopeless Dawn*, Boecklin's *Toteninsel*, Millais's *Boyhood of Raleigh*, Yeames' *When Did You Last See Your Father?*, Wallis's *Death of Chatterton*, Newcome's *What Shall He Be When Grown Up?*, Dali's *Composition*, an illustration from the *Strand Magazine* entitled *She Was Looking at Me through the Half Open Door*, and two drawings from *Punch*. These twelve pictures were used as a standard series because they were readily accessible to others who wished to use the test. In later experiments he found that better results could be obtained by using specially designed pictures suggested

¹ Steinthal, H. (1881), *Einleitung in die Psychologie und Sprachwissenschaft*, pp. 167f. The whole passage, quoted by James, *Principles*, II, p. 108, and James' own discussion are well worth reading. James contends that apperception is merely a pretentious and somewhat ambiguous name for "the sum total of the effects of association." He would prefer G. H. Lewes' word 'assimilation' (*Problems of Life and Mind*, 1879, I, pp. 118f., a discussion which emphasizes, with instructive illustrations, the individual nature of 'assimilations'). It was apparently James' objections to the term that caused its temporary disappearance from most American text-books.

by previous work and drawn by Miss Charleston ((16), p. 113). He describes the whole procedure as "resembling what older writers would have regarded as a 'free association' test, with concrete scenes for stimuli instead of words": "the visible form," he explains, "appeals more directly to primitive emotions and less to the higher critical feelings . . . whereas words lead more rapidly to emotional repression" ((10), pp. 326, 330).

The method of administration resembled that prescribed by Whipple for tests of description and report ((6), pp. 287, 301). In the group-form of the tests each child had a copy of the picture in front of him, and was asked first to write a free composition ('story') and then to answer an 'interrogatory.' Titles were erased, and had to be supplied by the child. With younger children a roneo'd outline of a story was presented to be filled up like a completion test. In the individual form of the tests the story and the replies were given orally. The schedule for marking the results, however, instead of referring solely to 'accuracy and range of report' (as in Whipple's version), included a set of headings for classifying and assessing replies, based on much the same scheme of personality as was used for other tests or assessments, noting (as had been done for the verbal association tests) the common types of clue to be found in the content of the responses and the apparent significance of each (21). The selection of clues was determined by a systematic item-analysis.

The quantitative results were only moderately successful: (average figures are, for reliability, 0.57, for validity, 0.16 to 0.38: (16), p. 116). But the method proved to be, like the association test, a useful starting point which could be followed up in private interviews on quasi-psychoanalytic lines. It was therefore, as he explains, regularly used by himself and his co-workers in studies of delinquent and neurotic children.¹

With the revival of associationism, owing to the influence of the behaviourist school,² interest in such concepts and procedures seems to have suffered an eclipse. There is no mention of apperception tests among the 'projection tests' described in the first edition of Cattell's *Guide to Mental Testing*, nor yet in Earl's excellent review of 'methods of assessing temperament and personality.'³ However, since the war the work of Murray and his colleagues at Harvard has led to widespread revival of interest in this type of test. To begin with, they apparently envisaged the method as a means merely for 'investigating fantasies' (11), but later realized it could be used for more general purposes (12).

The type of material and method of interpretation developed by these writers is excellently described by Vernon (22), who also gives a critical and impartial summary of the chief results. "Binet, Burt, and others," he writes, "have used picture interpretation as a means of studying intellectual and emotional qualities, but the test in its present form was first described by Morgan and Murray: they collected a standard series of 30 photographs; the subject is shown each picture, and asked to make up a story describing the situation, the events leading up to it, and the outcome, together with the thoughts and feelings of the characters. . . . Unfortunately the interpretation varies widely with the theoretical background of the psychologist. Those who follow Murray have adopted his elaborate system of needs and presses (external forces that the individual regards as beneficial or harmful): Rapaport and others present simpler schemes. They are based on the style or structure—compliance with instructions, language characteristics, logical coherence, consistency of stories with one another, realism, etc., and on the content of recurrent themes. . . . Burt and Sen (in an unpublished memorandum) advocate a more straightforward approach, owing little or nothing to abnormal psychology.

"Many writers have testified to the value of T.A.T. in mental hospitals, clinics, etc., especially when used with other tests such as Rorschach. But scientific evidence is conflicting. . . . Harrison was able to assess I.Q.s with a validity of .78. Sen finds the reliability of assessments by different scorers for Burt's qualities to average only .4; nevertheless such qualities as Observation, Verbal, Level of Organization, and Maturity gave very promising correlations with follow-up results among high-grade civil servants" ((22), pp. 181-4).

¹ Cf. (9), (10), and *The Young Delinquent* (1925), pp. 404f.

² Even to-day, if a student, curious to know what is meant by an 'apperception test,' turns to the index of his text-book of psychology, he will find the word neither in Woodworth nor in Boring, Langfeld, and Weld. The temporary disappearance of the term in British literature is due largely to Spearman's influence. Referring to "apperception and assimilation," he maintains that his doctrine of the education of relations and correlates (really a revised form of associationism) "shows that neither term can supply anything that psychology needs" (*The Nature of Intelligence and the Principles of Cognition*, 1927, p. 26).

³ Earl awards high praise to the inkblot test (as popularized by Rorschach and his followers)—a test which older writers showed was so largely dependent on apperception as above defined (14). Yet the intuitive methods of scoring the replies, as he describes, could only be accepted as plausible by one who was entirely unfamiliar with the theoretical and experimental work on apperception.

A Preliminary Study of the Thematic Apperception Test

Several correspondents have since asked if fuller details could be made available in regard to both the method of approach here referred to and the results obtained. Accordingly, I have endeavoured to summarize as succinctly as possible the general principles adopted in using and marking the test, and the main conclusions so far emerging in what (it should be emphasized) are merely preliminary trials.¹

II. SUBJECTS

The present investigation was based on the performances of 100 candidates for the Administrative Class of the Civil Service.

During the period of post-war reconstruction a new procedure was introduced by the First Commissioner (Sir Percival Waterfield), who also asked Sir Cyril Burt to act as Psychological Consultant to the Service and as a member of the Technical Advisory Committee. A full account of the general plan as finally adopted and of the tests and other methods employed has already been published by Mr. N. A. B. Wilson (19), who was at that time Principal Psychologist. Briefly it included a written qualifying examination, followed by a two- or three-day examination during a kind of 'house party' by a Civil Service Selection Board (which included psychologists and was known as CISSB), and ending with an interview by a Final Selection Board (FSB). A mark or grading was given at each stage, but the ultimate decision rested with the Final Selection Board. This Board judged each applicant on his entire record, including the marks received in the earlier stages, and his prowess at the final interview (20). Detailed results indicating the apparent value of this procedure and of its component parts (excepting the projection tests) have been given by Professor Vernon (22).

The main criteria used in the present research were derived from reports sent in on successful candidates a year or more after their appointment. The reports were submitted on special 'follow-up forms' containing spaces for (a) five-point ratings on the thirteen qualities believed to be most relevant to their duties, (b) an overall grading for general efficiency up to date, (c) an estimate of the highest rank in which the officer might be successful (termed the 'final rank'), and (d) a free pen portrait.

III. TEST MATERIAL AND MODE OF SCORING

Test Material. Neither the material used by Murray for American adults nor that used by Burt for school children would have been altogether suitable for candidates of this type. The pictures ultimately employed were the following: 1. A young woman surprised by an unexpected visitor; 2. A boy holding a violin, with a girl standing by him; 3. A man and a woman in conversation at a table; 4. Two puppets; 5. A woman sitting at a window, open book in hand, looking up at a man standing by her. Two-and-a-half minutes were allowed for the candidates to write a story about each.

At the time of my enquiry this set of pictures had not been in use long enough for a detailed item-analysis to be made of the various clues. The scheme of marking had consequently to be taken over very largely from work with somewhat different sets of pictures and a somewhat different type of examinee. Hence the whole enquiry can only be regarded as a tentative study of the applicability of the method proposed in a rather novel field.

¹ Mrs. M. M. McArthur (Principal Psychologist at CISSB) has very kindly read my draft; and I am greatly indebted to her both for her co-operation in the experimental work and for her criticisms of my original report. I have also to express my thanks to the Civil Service Commission for permission to publish results obtained while I was working for the Selection Board. They and their officers, of course, have no responsibility for any of the views here put forward or for the conclusions reached.

Mode of Scoring. As Rapaport observes ((17), p. 423), the usual techniques for interpreting the thematic apperception test "are not hard-and-fast rules like those of scoring other tests; they are rather viewpoints for looking on the stories, which have to become ingrained in the examiner." Murray and his collaborators drew up an elaborate scheme of 'needs' and 'presses' (which, as Cattell observes, is, in effect, an amalgam of McDougall's primary emotions with Freud's theory of fantasy formation); and "suggested and used depth interpretations which they compared with psychoanalytic interpretations of dreams" ((18), p. 214).

The immense number of headings involved in Murray's scheme of itself renders his method quite impracticable in any ordinary selection procedure; and the peculiar terminology employed—'conservance,' 'infavoidance,' 'similance,' 'succorance,' and the like—is exceedingly difficult to interpret with any exactitude. Those who have endeavoured to improve upon it still choose their qualities or traits in the light of intuitive speculation or *ad hoc* improvisation, and so produce widely varying schemes for which there is really very little empirical support. Full details of the methods and of the modifications suggested by their followers are given by Rapaport and Bell, and summarized by Vernon. Vernon repeatedly stresses the extreme subjectivity of all such procedures; and rightly adds that "the lack of any generally accepted framework for the description of personality creates further difficulties" ((22), p. 205). In addition it is hard to discover any precise information regarding either the reliability or the validity of the methods used by earlier investigators.

During their work with school children, extending over more than forty years, Burt and his co-workers have evolved a structural scheme for the study of personality; and in their researches on pictorial tests they have tentatively worked out a systematic procedure for analysing both the formal and material characteristics of the narrative compositions of children (9). It has been plainly influenced by the general practice of literary and biographical critics; but differs in retaining the same scheme for all the writers to be studied, in attempting to give a quantitative estimate for each characteristic, and above all in having been subjected to some degree of empirical verification. It employs a stock schedule for the classification of the commoner *motifs* (recurrent topics, subjects, or themes) according to the dominant emotions and salient interests (sentiments and complexes) which each appears to indicate, with special reference to their dynamic interactions.

The choice of traits has been based on the results of factorial studies ((8), (13), (16)), and checked by systematic comparisons with the after-histories of the cases tested; the inferences drawn can thus be fitted into a recognized psychographic scheme available for other tests. It is often objected that assessments for qualities derived by factorizing tests or traits fail to give a true synthetic portrait of the individual as a whole. To overcome this difficulty, Burt also grades the traits in order of their prominence in one and the same person. This enables a factorial 'psychogram' to be constructed, and yields a sounder estimate for what are sometimes called 'temperamental types' and should perhaps rather be termed 'temperamental tendencies.'¹

With a given set of pictures the commoner features of the stories which they tend to elicit are arranged according to the various cognitive or orectic characteristics they are presumed to indicate. In marking the scripts, a prepared schedule is generally used: pluses are entered in the appropriate spaces for the several signs or symptoms noted in each candidate's script (pluses or minuses in the case of bipolar traits). An allowance is given for over-all impressions. These scores are then added up, and treated as measurements for the several traits.

No doubt with pupils from the elementary schools the structure of the individual's personality is much simpler than that of the highly intelligent adult. Nevertheless, it seemed profitable to enquire whether this scheme might not provide a better starting-point than the unverified speculations which seemed to form the basis of the American techniques. In the investigations described below, an attempt was made to score the results of the test in accordance with these principles. The scheme covers both cognitive abilities and orectic (affective and conative) characteristics: but for practical reasons it could not be adopted in its entirety. In the present enquiry the figures quoted will be confined to those particular characteristics which

¹ Cf. (13), and *Factors of the Mind*, p. 427, Fig. 1.

seemed to be most closely related to the duties that would be required of the successful candidates and which previous attempts had shown could be rated with greatest reliability.

IV. RELIABILITY AND VALIDITY OF ASSESSMENTS

A. Reliability. The stories written by a hundred candidates were scored by Mrs. McArthur and myself. Further, the scripts of twenty candidates at an earlier examination, which I had already scored in the same way, were re-marked after an interval of nine months, when, owing to intervening work, I found I had completely forgotten my original impressions. The results obtained (*a*) for the self-consistency of the same interpreter and (*b*) for the reliability of the two independent assessments are shown below in Table I. With a sample of 100, all product-moment coefficients over 0.197 and 0.256 are significant at the 5 per cent. and the 1 per cent. levels respectively; with a sample of 20, coefficients above 0.444 and 0.561 are significant at those levels.

TABLE I. COEFFICIENTS OF SELF-CONSISTENCY AND RELIABILITY

Traits Assessed				Self-consistency of One Judge	Reliability of Two Judges
<i>A. Cognitive</i>					
1.	General Organization of Story	..	..	(assessed only once)	
2.	Observation	..	..	.492	.287
3.	Verbal Ability	..	..	.758	.467
4.	Imagination	..	..	.605	.588
<i>B. Orectic</i>					
5.	General Emotionality	..	..	.634	.395
6.	Extraversion-Introversion	..	..	—	.222
7.	Cheerful-Depressive	..	..	.715	.628
8.	Self-assertive-Self-submissive	..	..	.426	.466
9.	Anxiety	..	..	—	.451
10.	Decisiveness	..	..	.560	.418
11.	Ambition	..	..	.169	.295
12.	Social Consciousness	..	..	.804	.452
13.	Maturity-Immaturity	..	..	.718	.485
14.	Integration	..	..	—	.371
Average				.588	.425

If one may judge from the more careful studies reported in the literature, the reliability of so-called 'projective tests of personality' seldom reaches a figure of 0.400; and this was approximately the average value in our previous attempts. The table above shows that all except one of the self-consistency coefficients, and nine of the reliability coefficients, rise above that level. Somewhat unexpectedly the quality which is most consistently assessed is Cheerfulness. The three poorest reliabilities which fall well below the figure of 0.400 are those for Observation, Ambition, and Extraversion-Introversion: the last indeed is non-significant. Further study indicated that, in these cases, the low coefficients were due to the fact that the two investigators seemed often to be looking for slightly different things: e.g., in assessing Observation, one of the judges looked for details regarding physical appearance, dress, gesture of the characters in the story (i.e., indications of observation in general life), whereas the other interpreted the term more narrowly as meaning accurate observation of what actually appeared in the picture. In this and other instances much of the discrepancy arose from an inadequate definition of the quality to be assessed and from incomplete

specifications of the relevant clues. The term *Ambition*, for example (described as 'an ardent desire for personal distinction or glory'), is a somewhat ambiguous concept; and the phrase 'Self-assertion *versus* Self-submission' (described as McDougall defines those tendencies) appears to treat aggressive tendencies and self-assertive tendencies as almost synonymous: although correlated, it seems clear that the two should be assessed separately. Moreover, some of the pictures were not altogether suited to elicit trustworthy clues to the qualities required. But these are defects in the procedure which could doubtless be corrected after further research.

B. *Validity*. In judging validity the criteria used were (a) the marks allotted by the Final Selection Board after the three-day examination and (b) the various marks and gradings provided by the follow-up forms. Data sufficiently complete for comparison were available for only 42 of the successful candidates. The scores for all the traits were first correlated with suitability as estimated (i) by the Final Selection Board, (ii) by the Over-all Grading, and (iii) by the estimate of 'Final Rank.' Table II shows the results.

With samples of this size, correlations above 0.304 and 0.393 are significant at the 5 per cent. and 1 per cent. levels respectively. The correlations are no doubt greatly reduced by the fact that those candidates who obtained low marks for desirable qualities failed to secure appointment; they had consequently to be omitted from the calculations for validity. Any attempt to correct for this type of selection by statistical calculation seemed highly precarious. Hence the figures are printed as they stand, and must be compared with uncorrected coefficients given in other reports, not with coefficients already corrected for selection.

TABLE II. CORRELATIONS WITH SUITABILITY

Traits Assessed		(i) Final Selection Board	(ii) Over-all Grading	(iii) Final Rank	Average
A. <i>Cognitive</i>					
1.	General Organization of Story ..	.427	.318	.354	.367
2.	Observation	.268	.331	.401	.333
3.	Verbal Ability	.344	.237	.359	.313
4.	Imagination	.328	.168	.254	.250
B. <i>Orectic</i>					
5.	General Emotionality	.127	.100	.203	.143
6.	Extraversion-Introversion	.219	.118	.080	.108
7.	Cheerful-Depressive	.117	.147	.183	.149
8.	Self-assertive-Self-submissive	-.070	-.228	-.151	-.150
9.	Anxiety	-.149	-.130	-.104	-.128
10.	Decisiveness	.198	.074	.137	.136
11.	Ambition	.029	.044	.027	.033
12.	Social Consciousness	.037	.094	.207	.113
13.	Maturity-Immaturity	.309	.183	.335	.276
14.	Integration	.233	.299	.325	.286
Average		.204	.170	.231	.201

NOTE. In Tables II and III the negative signs have been reversed in calculating the averages at the foot of the columns.

Owing to the small size of the sample, only six traits furnish one or more correlations that are statistically significant, and of these four are cognitive. Nevertheless, the figures are on the whole higher than the validities obtained for most of the selection traits used at CISSB (4), where the highest (uncorrected) correlation with the criterion was 0.186 (for the 'Short Talk'). In general the estimate

A Preliminary Study of the Thematic Apperception Test

of Final Rank correlates more closely than the other two. The highest average correlation is obtained by the marks for 'General Organization of Story.' This is the only characteristic that provides a fully significant coefficient with all three gradings: it may perhaps be taken as a measure of the kind of verbal and logical intelligence that is required in such candidates. The Final Selection Board mark correlates more highly than the other estimates with the scores for Organization of Story, Imagination, Lack of Anxiety, Extraversion, and Decisiveness. The estimate of Final Rank, which gives a long-term view of a candidate's suitability, correlates most highly with Observation and Verbal ability among the cognitive qualities, and with Maturity, Integration of Personality, Social Consciousness, Emotionality, and Cheerfulness among the orectic. The former group of qualities are in the main those which we might perhaps expect to tell most at a personal interview; the latter, those which on the whole might be expected to make for general efficiency of work in a government department.¹

Since different Government departments require slightly different qualities, an attempt was made to determine how far these more specific qualities, as assessed by the supervisors, could be predicted by the marks for the several traits that approximate most closely to these qualities. Table III shows the correlations between the most relevant assessments derived from the test and the nearest assessments appearing on the report forms. All the correlations point in the right direction. But the only correlations that are significant at the 5 per cent. level are those for Observation and Imagination.

TABLE III. CORRELATIONS FOR SPECIFIC TRAITS

(a) Traits Assessed by Test		(b) Traits Assessed on Follow-up Form	Correlation
2.	Observation	Attention to detail	.304
3.	Verbal Ability	Paperwork (minutes and correspondence) ..	.256
4.	Imagination	Fertility of ideas	.384
6.	Extraversion-Introversion	New contacts	.281
6.	Extraversion-Introversion	Acceptability as colleague	.116
7.	Cheerful-Depressive	Judgement	.196
8.	Self-assertive-Self-submissive	Acceptability as colleague	-.260
8.	Self-assertive-Self-submissive	Management of subordinate staff	-.204
10.	Decisiveness	Judgement	.264
		Average	.252

So far as the foregoing results take us, it may, I think, be reasonably claimed that the analytic method here adopted for evaluating the results obtained with a pictorial apperception test is unquestionably the most satisfactory for securing objective evidence regarding the reliability and validity of such procedures. It would be instructive to compare the over-all picture that might be derived for each candidate from the detailed assessments with the results of the synthetic judgements submitted by those who prefer to rely on intuitive impressions. What, however, seems most urgently needed is a further improvement of the details of the technique. Probably it would be wiser to begin first by selecting a more informative set of pictures in the light of the experience already gained. After that a systematic item-analysis should be undertaken to determine first those traits that can be most reliably assessed by this particular method and then what detailed clues obtained with the pictures selected yield the most valid judgements.

¹ A factor analysis of the follow-up form (20) showed that the Over-all Grading falls on the same side of the bipolar analysis as the cognitive qualities (such as knowledge of work, fertility of ideas, quick apprehension, and attention to detail), whereas Final Rank is grouped with the orectic or 'personal' qualities (long-term contacts, new contacts, and effectiveness as member of a committee).

V. SUMMARY AND CONCLUSIONS

1. The responses to a thematic apperception test obtained from a hundred Civil Service candidates in the course of the regular entrance examination were scored by two investigators according to Burt's method of rating for a systematic scheme of traits based primarily on the results of earlier factorial analyses. The correlations between separate assessments made by the same judge at an interval of nine months showed a moderately high consistency for nearly all the traits (average 0.59). The correlations between the independent ratings made by two judges showed that, when compared with similar coefficients for other projection tests of personality, or other methods of marking, the assessments obtained by this method possessed a degree of reliability that is encouraging, though still not high enough to have much practical value as they stand. Three traits at least seem to be ill-defined and possibly may not lend themselves to assessment by this type of test. The traits yielding the largest reliability coefficients were Verbal Ability, Imagination, Cheerful or Depressive tendencies, Self-assertive or Self-submissive tendencies, Anxiety, Decisiveness, Social Consciousness, and Maturity versus Immaturity.

2. The validity coefficients are higher than those obtained with other tests of the same characteristics and other methods of marking, and are superior to those hitherto published for similar assessments of Civil Service candidates: the correlations between the assessments based on the test and the various criteria available suggested that, of the traits assessed, those most closely related to suitability for administrative work appeared to be Ability for Coherent Organization of the Story, Observation, Verbal Ability, Imagination, Maturity, Cheerfulness, and Integration of Personality. Correlations with reports on the successful candidates' efficiency sent in a year or so after their appointment indicated that several of the more specific traits could be assessed with a promising degree of accuracy.

3. The enquiry as a whole indicates that both the schedule of traits and the method of scoring form a distinct improvement on those adopted by previous British workers (for the most part borrowed, with some simplification, from American contributions). At the same time the detailed results reveal that certain features in the procedure leave considerable room for improvement; and a study of the chief sources of discrepancy has suggested what will probably be the most fruitful directions for future research. It seems clear that the efficacy of the thematic apperception test can be greatly improved if the traits to be assessed are based on an acceptable theory of personality-structure, and if the various clues used in the marking-scheme are selected on the basis of a systematic item-analysis and are explicitly specified and defined.

REFERENCES

1. Bryant, S. (1886). 'Experiments in testing the character of school children.' *J. Anthropol. Inst.*, XV, 338-49.
2. Stout, G. F. (1888). 'The Herbartian Psychology.' *Mind*, XIII, 321-38, 473-98; XIV, 1-26, 353-68.
3. Binet, A., and Henri, V. (1896). 'Psychologie individuelle.' *Ann. Psychol.*, III, 296-332.
4. Binet, A., and Simon, T. (1908). 'Le développement de l'intelligence chez les enfants.' *Ann. Psychol.*, XIV, 1-94.
5. Adams, J. (1910). *The Herbartian Psychology applied to Education*. London: Heath.
6. Whipple, G. M. (1910). *Manual of Mental and Physical Tests*. Baltimore: Warwick and York.
7. Burt, C., and Moore, R. C. (1912). 'The Mental Differences between the Sexes.' *J. Exp. Ped.*, I, 273-84, 355-88.

A Preliminary Study of the Thematic Apperception Test

8. Burt, C. (1915). 'General and Specific Factors underlying the Primary Emotions.' *Brit. Ass. Ann. Rep.*, LXXXIV, 694-6.
9. Burt, C. (1934). 'The Interpretation of Pictures.' Unpublished Laboratory Notes. University College, London.
10. Burt, C. (1935). *The Subnormal Mind*. London : Oxford University Press.
11. Morgan, C. D., and Murray, H. A. (1935). 'A Method for Investigating Fantasies : the Thematic Apperception Test.' *Arch. Neurol. & Psychiat.*, XXXIV, 289-306.
12. Murray, H. A., et al. (1938). *Explorations in Personality*. New York : Oxford University Press.
13. Burt, C. (1939). 'The factorial analysis of emotional traits.' *Character and Personality*, VII, 1-238, 285-99.
14. Earl, C. J. C. (1939). 'Methods of assessing temperament and personality,' ap. Bartlett, F. C., et al. *The Study of Society* (ch. X). London : Kegan Paul.
15. Burt, C. (1945). 'The Thematic Apperception Test.' Unpublished Laboratory Notes. University College, London.
16. Burt, C. (1945). 'The Assessment of Personality.' *Brit. J. Educ. Psychol.*, XV, 105-21.
17. Rapaport, D. (1946). *Diagnostic Psychological Testing*. Chicago : World Book Company.
18. Bell, J. E. (1948). *Projective Techniques*. New York : Longmans, Green & Co.
19. Wilson, N. A. B. (1948). 'The Work of the Civil Service Selection Board.' *Occup. Psychol.*, XXII, 204-12.
20. Vernon, P. E. (1950). 'The Validation of Civil Service Selection Board Procedure.' *Occup. Psychol.*, XXIV, 75-95.
21. Sen, A. (1950). 'A statistical study of the Rorschach test.' *Brit. J. Psych., Stat. Sect.*, III, 21-39.
22. Vernon, P. E. (1953). *Personality Tests and Assessments*. London : University of London Press.

THE FACTOR ANALYSIS OF MATRICES OF NEGATIVE CORRELATIONS

By PATRICK SLATER
Bedford College, University of London

I. *Introduction.* II. *Statistical Conditions for Negative Matrices.* III. *Experimental Conditions.* IV. *Analysis of New Example.* V. *Analysis of the Previous Example.* VI. *Methodological Discussion.* VII. *Summary.*

I. INTRODUCTION

In the discussion of a matrix of predominantly negative correlations in a previous paper (1), a first attempt was made to define the general conditions under which such matrices might occur. Taking into consideration another instance of a slightly different type (cf. Sect. III), a more adequate definition can now be reached.

In the earlier paper only a few incidental comments were made on the possibility of applying factor analysis to such matrices. It was mentioned that an imaginary general factor (i.e., one with coefficients of $i = \sqrt{-1}$ for its loadings) might be obtained, if a centroid analysis were applied mechanically, and that a plausible psychological explanation might even be given for such a factor, since a particular range of mental processes (involving an element of choice) would be found in most psychological experiments where negative matrices were likely to occur. But it was shown that the need for extracting an imaginary general factor could be eliminated by introducing a transformation.

Burt, in a *Reply* in the same issue (2), expressed the view that a factor analysis would be possible without the extraction of any imaginary factor, and he doubted whether any preliminary transformation was required. He gave a factor analysis of my figures, but did not include a complete answer to the question discussed in detail here, how factor analysis should be applied to such matrices. However, the main conclusion now reached is in close agreement with Burt's original opinion. It is in fact shown that an efficient analysis is possible without transforming the matrix or extracting any imaginary factor, a transformation being only required in the test of significance.

Both the original example and the one adduced here are instances of matrices of variances and covariances of variables in a common metric, which do not need to be standardized. This permits the use of a variant of Bartlett's test ((3), (4), and (5)) which has theoretical advantages but is appropriate only when factor analysis is applied without prior standardization. In experiments involving ratings, rankings, and comparisons, where negative matrices are particularly likely to occur, a common metric can frequently be used for recording the simultaneous variations; and an illustration of the treatment of unstandardized matrices may be found helpful.

II. STATISTICAL CONDITIONS FOR NEGATIVE MATRICES

Suppose a matrix of correlations has been obtained by measuring m variables, A, B , etc., in n cases, the observed measurements in a particular case being $a_0, b_0 \dots m_0$. The sum of these measurements, s_0 say, will usually also vary from case to case, and may be treated as a further variable, S . If V_a, V_b , etc., denote the total variances of the variables and W_{ab} , etc., their covariances, we may write $\Sigma V = V_a + V_b + \dots V_m$, and $\Sigma W = W_{ab} + \dots W_{lm}$ (summing the covariances without duplication) and shall then obtain, for the total variance of S , $V_s = \Sigma V + 2\Sigma W$, i.e., the sum of all the terms in the square matrix of total variances and covariances. When ΣW is negative, i.e., when $\Sigma V > V_s$, the matrix of correlations will be preponderantly, though not necessarily completely, negative.

This may happen if the conditions of the experiment restrict the variation of S directly, while only indirectly restricting A, B , etc. If S is completely restricted, $V_s = 0$, and consequently $\Sigma W = -\frac{1}{2}\Sigma V$. This is evidently the negative limit of ΣW when A, B , etc., are real scalar measurements. For if it exceeds this, V_s must be negative; but as V_s is a sum of squares, it cannot be negative unless some of its component terms, such as s_0 , are imaginary or complex; and as s_0 too is a sum, this implies that some of its components, a_0 , etc., must be imaginary or complex also. In short, some at least of the dimensions in which variation has been recorded must have been imaginary or complex *ab initio*. Matrices of this kind are not discussed below, for lack of any experimental example. But there would not appear to be any objections in principle to extracting imaginary factors from them.

Where $V_s = 0$, the conditions of the experiment must have excluded variation in S ; and thus the matrices obtained resemble matrices from which a variable has been partialled out statistically. Methods for their analysis can be adapted from methods appropriate for bipolar matrices of residual correlations. Difficulties in their treatment arise because their determinants are zero, and because in the absence of association the correlation between pairs of variables does not tend to 0 but to $-1/(m-1)$. Tests of significance are complicated accordingly. Degrees of freedom and numbers of latent roots also require special attention.

When V_s is positive but less than ΣV , the matrix may be unusual in superficial appearance only; but the difficulties which are encountered when $V_s = 0$ must become gradually more obtrusive as it approaches this limit. The value of the determinant must recede towards vanishing point, and the null value of the correlation coefficient shift from 0 towards $-1/(m-1)$.

The method of principal components can be applied to either type of matrix without postulating any imaginary factor. When all the observed variance is in real scalar dimensions it would be difficult to find valid grounds for inferring the occurrence of variation in any imaginary dimension. And the extraction of an imaginary factor would leave the residual variances of the variables greater than their original totals. Generally, therefore, to postulate such an imaginary factor would appear unnecessary, improper, and absurd.

III. EXPERIMENTAL CONDITIONS

A simple instance of a negative matrix resulting from a partial restriction of V_s is provided by the following data from an investigation into the personal histories of two groups of men in training for an operational duty. One group consists of men already in the service posted on to the course from their previous units; the other of men who have re-entered as volunteers after completing a previous period of service and returning to civil life for a time.

The number of years they had completed since reaching the age of 14 in (E) full time education (outside the service), (C) civilian employment, and (F) service (including technical training) were recorded, these being the only ways in which substantial parts of their lives had been spent. Since

the age for admission to the duty is restricted, those who have occupied more time in one way will have passed less in another; and the three variables are negatively correlated. Table I gives the main statistics.

TABLE I. ANALYSIS OF DATA FROM MEN IN THE FORCES

Variable	A. Means		B. Mean Square			C. Correlations		
	Trans- ferred (<i>n</i> = 54)	Re- entrants (<i>n</i> = 16)	Variances and Covariances (<i>d.f.</i> = 69)			E.	C.	F.
(E) Education	1.65	2.25	1.5912	-.9213	-1.3582	—	-.3340	-.4078
(C) Civil employment	1.39	4.19	-.9213	4.7818	-.6431	-.3340	—	-.1114
(F) Service	5.13	3.19	-1.3582	-.6431	6.9723	-.4078	-.1114	—

When the simultaneous variation of a number of variables is subject to the condition that some cannot increase unless others diminish, a choice between the variables must occur at some point. This may not always be the deliberate choice of a human being, but in psychological experiments it often is. Thus in this experiment an element of choice may be found in the way in which each man decided to employ his time: the statistics represent the accumulated effects of such acts of decision. It may also be found in the alternatives open to the selection board, who cannot simultaneously require extensive experience in many different fields if they wish to recruit men still in the prime of life.

In psychological experiments with methods of ranking or paired comparison the entire variation is usually attributable to acts of deliberate choice. In preferring one of the alternatives presented to him the subject declines another. Correlating preferences for different alternatives thus gives rise to predominantly negative matrices. Face to face with death there is not much deliberate choice open to an individual, but the situation is analogous in that by dying in one way he escapes dying in any other. A choice is made for him if not by him. Thus negative correlations are commonly found when fluctuations in mortality from different causes are correlated over a period during which the numbers and age composition of the population are tending to remain constant.

If such analogies are admissible, it would seem reasonable to enquire after an element of choice in the experimental conditions whenever a negative matrix is observed. Conversely, when an experiment involves some such element, some approximation to a negative form of matrix may be expected, and the method of analysis adopted should take this into account.

The mortality statistics illustrate another point. If the period under consideration is one when the numbers and age composition of the population are undergoing substantial changes, years when deaths from any particular cause are common are likely to be years when the total number of deaths is large; and a general tendency towards positive correlation may appear among the causes. In the former case the total number of deaths from all causes (*S* in the notation of Sect. II) is not free to vary widely; in the latter it is. The transition from predominantly positive correlation matrices, to vary widely; in the latter it is. The transition from predominantly positive correlation matrices, to predominantly negative ones, productive yielding positive general factors of the familiar kind, to predominantly negative ones, productive (when improperly analysed) of imaginary general factors, may thus be seen to be gradual and subject to control by any processes or conditions which expand or contract V , relatively to ΣV .

IV. ANALYSIS OF NEW EXAMPLE

When a three-dimensional model of the total dispersion matrix in Table I B was constructed, no startling peculiarities were found. Its axes, computed by the method of principal components, were found to be as shown overleaf. Since the variables were expressed in a common unit (number of completed years) they were left unstandardized.

The results may be compared with those obtained by a discriminant analysis. The function maximizing the discrimination between the two groups requires the relative weights: (E) -1.3465, (C) -1.3151, (F) +.1270; 46.25 per cent. of its total variance is attributable to the differences between the groups. The determinants of the matrices of variances and covariances, total and within groups, have the corresponding ratio, 11839632 : 6365674.

Variable	Factor loading		
	First	Second	Third
(E) Education	.. — .5422	.. — .6668	.. .9234
(C) Civilian employment	.. — .4606	.. 2.1213	.. .2638
(F) Service	.. 2.6193	.. .2360	.. .2369
Latent root	7.3667	5.0004	.9783

In internal factor analysis it is generally assumed that if the individuals measured differ in some respect, and the measurements recorded are sensitive to the difference, the major dimensions along which the individuals are found to vary may be used to infer and define the nature of the difference between them. This is the basic factorial hypothesis popularly translated into a belief that factor analysis can be used to prove or disprove the existence of psychological traits.

The major axis of the total dispersion matrix should accordingly approximate to the axis along which the dispersion between the groups is greatest. Actually the two axes diverge considerably. The major amount of variation between the individuals is in their length of service (cf. the first factor, above) which involves, as it increases, a diminishing length of time spent on education or in civilian employment. The variation between the groups, however, is mainly due to differences in the amount of time they have engaged in full time education or civilian employment, especially the former (the discriminant weights need to be related to the mean square variances, Table I B). Pooling square methods may be used to show that the two axes correlate by .57, i.e., lie at an angle of approximately 55° to one another. The difference between the groups contributes only 14.6 per cent. of the total variance of the first component, and slightly more, 18.1 per cent., of the second.

V. ANALYSIS OF THE PREVIOUS EXAMPLE

The data given in the previous paper were obtained by asking five groups, saleswomen (S), light factory workers (F), teachers (T), nurses (N), and clerks (C), 18 women in each group, to score the five occupations for smartness of appearance off duty by the method of paired comparisons. Here s_0 , the sum of the scores assigned by each subject, was constant and consequently $V_s = 0$.

The matrix of mean square variances and covariances is shown in Table II A below (obtained from the sum of the two matrices given in Tables III A and III B of the original paper). The principal components are given in Table II B.

TABLE II. ANALYSIS OF DATA FROM EMPLOYED WOMEN

Variable	A. Variances and Covariances (d.f. = 89)					B. Principal Components			
	S.	F.	T.	N.	C.	First	Factor loading Second	Third	Fourth
S ..	5.0164	.8904	—3.7686	—1.8382	— .3000	—1.9044	.4136	— .2329	1.0791
F ..	—	3.2265	—2.2000	—1.6427	— .2742	—1.0779	— .2852	1.2196	— .7041
T ..	—	—	6.0270	.7034	— .7618	2.1845	— .7602	.3727	.7336
N ..	—	—	—	4.2517	—1.4742	1.0983	1.6652	— .2513	— .4576
C ..	—	—	—	—	2.8102	— .3005	—1.0334	—1.1081	— .6512
Latent root	—	—	—	—	—	10.8572	4.7613	2.9716	2.8319

Teaching is associated with nursing at one end of the major axis of variation, and sales with factory work at the other; the diversity of opinion about the relative status of the professions is familiar and may be treated as a factor. The first two factors have loadings comparable with those

obtained by Burt (2) with a different procedure. The analysis concludes with the extraction of four factors, the fifth having vanished.

The two major components (λ_1 and λ_2) coincide closely with the two principal axes (μ_1 and μ_2) of variation between groups, previously published. Their inter-correlations are as follows:

	μ_1	μ_2
λ_1	-.9261	.2998
λ_2	.3733	.8055

The basic factorial hypothesis (that the difference admitted by the sampling will be located by the analysis) attains a much greater validity here.

VI. METHODOLOGICAL DISCUSSION

Factor analysis is usually applied in experiments in which variation has been recorded in a relatively large number of original dimensions, with the intention of showing that a smaller number of derived dimensions will account for most of it. The dimensions in which variation is greatest are found first (the method of principal components being the most efficient); and the remainder may be presumed to be dispersed by chance approximately equally among the other dimensions. Bartlett uses this hypothesis for his test of significance.

Without modifying the intention of the experiment, the introduction of the restriction described above modifies the outcome. As V_s diminishes towards 0, the dispersion contracts in one of its m dimensions, and finally collapses into a hyperplane of $m - 1$ dimensions. The latent root thus reduced recedes to last in order of extraction before ultimately disappearing. Thus in the first example the last latent root is relatively small; and in the second it has gone.

It seems doubtful whether Bartlett's test is suitable when V_s is partially restricted. As the significance of each successive component is judged by comparison with the remainder, a disproportionate reduction in one of the roots must enhance the apparent significance of its predecessors. When it has vanished, this difficulty disappears.

The test assumes its simplest and most general form when applied to a matrix of unstandardized variables, such as Table II. After k roots have been extracted the reduced χ^2 needed for testing the significance of the remainder is

$$\chi^2 = - \left\{ n - \frac{1}{6} (2p + 1 + 2/p) - \frac{2}{3} k \right\} \log_e V_{p-k},$$

with $\frac{1}{2} (p - k - 1) (p - k + 2)$ d.f. Here n is the d.f. of the variables, and p the total number of non-zero latent roots (i.e., $m - 1$ in the notation of Sect. II). V_{p-k} corresponds to R_{p-k} in (3), the latent roots of the unstandardized matrix being scaled down to an average of 1.

For the additive table of χ^2 in Table III the coefficient of $\log_e V_2$ has been used for all the values of V_{p-k} . The first latent root is the only one that appears to be beyond the limits of chance.

TABLE III. SIGNIFICANCE TESTS

Latent Root	Analysis of χ^2 d.f.		Values of V_{p-k}
λ_1 2.0273	4	48.90	V_4 .52884
λ_2 .8890	3	7.59	V_3 .91747
λ_3 .5549	2	.05	V_2 .99942
λ_4 .5288	—	—	—
Total	9	56.54	—

If the test is to be applied before all the latent roots have been extracted, their mean, needed to scale them down, can be found from $\Sigma V/p$, and their product, needed for V_p , from $|T'MT|$. Here M is the matrix to be analysed, ΣV is the sum of the terms in its leading diagonal, p the number of latent roots, and T the transformation matrix quoted in the previous paper, with transpose T' .

The Factor Analysis of Matrices of Negative Correlations

Such a transformation may thus be used for the test, but is not needed for the analysis. When the components are extracted directly from M , their loadings are related directly to the original dimensions, and are thus easier to interpret. The use of T is therefore not recommended except for the test.

If a disproportionately small latent root is suspected as a result of a restriction on V_s , it could be eliminated by partialling out the variance of S ; and the analysis would then follow the second model. To treat the matrix in Table I B in this way, chronological age would be held constant. But such treatment might not be suitable or even clearly intelligible in every instance.

The factorial hypothesis would not ordinarily be required when the origin of the heterogeneity in the sample is known independently, as in these examples. It is more likely to be helpful when it is known that the sample is heterogeneous in some respects but the individuals composing it cannot be classified independently. However, its application to such samples as these illustrates its validity, which is of course limited and varies from one experiment to another.

VII. SUMMARY

1. A matrix of negative correlations is defined here as one derived from exclusively or at least predominantly negative covariances. Such matrices can occur even when the variables concerned yield only real scalar measurements, provided the sum of the measurements in each case can vary only to a limited extent if at all. They are particularly likely to occur in experiments fixing a limited number of alternatives between which a choice has to be made or imposing an analogous set of conditions. The transition from a predominantly positive to a predominantly negative matrix occurs gradually as the limitations become more rigid.

2. To postulate an imaginary general factor in analysing them is unnecessary and confusing: their principal components can be found directly; no transformation is required. But the fact that as the transition from positive to negative occurs one of the latent roots is reduced and finally eliminated is important for factor analysis. If it is eliminated, a test of significance can be applied; and here a transformation may be useful. No direct solution is offered for the problem of testing significance when it is merely reduced, but it may be avoidable in some instances by partialling the root out completely before proceeding to analyse.

3. Circumstances in which factor analysis might be applied have been described; and certain problems involved in determining the psychological connotations of factors, and in assessing their validity, have been briefly discussed in connexion with certain concrete examples.

REFERENCES

1. Slater, Patrick (1951). 'The transformation of a matrix of negative correlations.' *Brit. J. Psych., Stat. Sect.*, IV, 9-18.
2. Burt, C. (1951). 'Reply to Mr. Slater's Criticisms.' *Brit. J. Psych., Stat. Sect.*, IV, 18-20.
3. Bartlett, M. S. (1950). 'Tests of significance in factor analysis.' *Brit. J. Psych., Stat. Sect.*, III, 77-85.
4. Bartlett, M. S. (1951). 'A further note on tests of significance in factor analysis.' *Brit. J. Psych., Stat. Sect.*, IV, 1-2.
5. Bartlett, M. S. (1951). 'The effect of standardization on a χ^2 approximation in factor analysis.' *Biometrika*, XXXVIII, 337-44.

NOTES AND CORRESPONDENCE

ON THE OBLIGATION TO REFUTE RIVAL HYPOTHESES

By L. F. RICHARDSON

In the recent controversy between Dr. H. J. Eysenck and his reviewer (W. L. G.), much emphasis is laid by the latter on the need to disprove alternative hypotheses (this *Journal*, V, pp. 208f.; VI, pp. 44f.). I am impressed both by the importance of this requirement and by the difficulty of satisfying it. The practical difficulty is obvious: for how can anybody be certain that he has thought of all the possible alternatives? They can be so extremely diverse. Take, for example, a long-lasting theory, that of Sir Isaac Newton about the motion of the planets. At the time when it was first discussed, permit me to imagine that there may have been people whose alternative was: "God created the planets; the planets obey their Creator; this view of their movement is comprehensive and satisfying; Mr. Newton's notions about attraction and momentum are superfluous." Even to-day it would be difficult to refute that alternative: all we can say is that it yields no predictions. In the next century Clairaut's alternative theory was that the attraction did not vary precisely as the inverse square, but as some slightly different function of the distance (H. Lamb, *Dynamics*, Cambridge, 1920, p. 271). From the year 1916 the much-discussed alternative theory was Einstein's, and it was strangely different from any alternative that had been proposed in the previous two centuries. Sir Harold Jeffreys, after mentioning some small discrepancies between theory and observation that still remain unexplained, says: "There has not been a single date in the history of the law of gravitation when a modern significance test would not have rejected all laws and left us with no law. Nevertheless the law did lead to improvement for centuries . . ." (*Theory of Probability*, 2nd ed., p. 362).

As a consequence of W. L. G.'s insistence, I try now to refute the alternatives to my theory of arms-races (this *Journal*, current number, p. 89). Very briefly my own theory is: "The present arms-race is due to mutual stimulation." What are the alternative theories? The theory prevalent in Britain and U.S.A. is: "The present arms-race is entirely caused by the aggressive policy of Russia and her satellites." Another theory, prevalent in Communist countries, is that "the present arms-race is entirely caused by the aggressive designs of the capitalist countries." I do not indeed feel able to refute these theories one at a time: but do they not refute each other, at least as far as the assertion 'entirely' is concerned? What is left of each, when 'entirely' is omitted, is the mutual stimulation which I have formulated.

REPLY BY THE REVIEWER

I am glad that so experienced a scientist as Dr. Richardson has taken up the questions raised by the recent enthusiasm of certain psychologists for a purely 'hypothetico-deductive methodology,' and that he agrees, at any rate in principle, with what I said about the need to refute alternative hypotheses. I willingly accept his main criticism of my argument, namely, that it failed to recognize the practical difficulties involved in trying to satisfy the logical requirements. If I did not explicitly refer to this point, that was partly because my reply to Dr. Eysenck was already too long, and partly because, in this portion of it, I was merely summarizing views already expounded at greater length by Burt in *The Factors of the Mind* (pp. 31-47).¹ May we hope that he himself will deal more fully with the questions raised?

Dr. Richardson, like Dr. Eysenck and several other writers, cites Newton's *Principia* as a model; and on this point I had, in a previous contribution (this *Journal*, V, p. 208), quoted and briefly endorsed Burt's remark deprecating the unfortunate way in which "earlier psychologists—Hume, Adam Smith, Bentham, and James Mill—all, explicitly or implicitly, took Newton's hypothetico-deductive method for their model." Obviously, however, I should have offered evidence for this view. There are several reasons why Newton's method seems inappropriate in psychology.

(1) It is generally agreed that in the *Principia* Newton deliberately decided on a hypothetico-deductive method of exposition, because he thought that would be the most convincing way to

¹ I am indebted both to Professor Burt and to Dr. Richardson for reading my note in draft, and for their written comments and corrections: Professor Burt has also allowed me to make use of his lecture-notes for references and illustrations. But for the applications made in reference to the present issue I alone am responsible. Incidentally, I am puzzled by the importance Burt attaches to 'Bacon's method': Jevons, de Morgan, and Welton all assure us that "it has not been followed by any of the great masters of science" (*Principles of Science*, p. 507).

present his theory to his own contemporaries: instead of "the new inductive method," he adopted the "classical geometrical method," which started with definitions and axioms (or "laws"), and then deduced a succession of theorems by a train of demonstrative reasoning. This would be the only orthodox procedure for the Cartesians, who would form his main critics.

(2) The phenomena with which Newton was concerned in the *Principia* were far simpler than those dealt with by the student of individual psychology. The number of variables involved—a few large bodies each possessing mass, and moving with the passage of time in three-dimensional space—was sufficiently small for a comparatively exact analysis to be made by relatively simple mathematical procedures. The most rigorous behaviourist would never attempt to "deduce" the orbit of a golfer hunting for his ball from three hypothetical laws of motion and one universal principle in the way Newton attempted to "deduce" the orbit of the moon.

(3) What Dr. Richardson calls Newton's "long lasting theory about the motion of the planets" is not a 'hypothesis,' in the sense in which I myself was using that highly ambiguous term. I was thinking rather of what Dr. Eysenck had called a 'hypothesis,' namely (in W. E. Johnson's phrase), "the initial formulation of a general proposition which it is proposed to establish." In the terminology adopted by Kneale (*Probability and Induction*, 1949, pp. 92f.), Newton's theory of universal gravitation is a 'transcendent hypothesis' to be established by 'secondary induction': we accept it chiefly because it co-ordinates a host of empirical generalizations by means of a simple and fruitful set of postulates. But what I had in mind was an 'empirical generalization' of comparatively narrow range, such as might be established by 'primary induction' as the result of a single factorial research.

(4) Nevertheless, it is not quite right to suppose that Newton considered it unimportant to disprove alternatives. Some had already been disproved by his predecessors. For example, Galileo's investigation of the problem of falling bodies proceeds in this way: indeed it provides the first instance of what has been called the hypothetico-disjunctive method—that combination of logical reasoning and planned experiment which is so distinctive of modern inductive science. Everyone knew that a ball rolling downhill would travel with increasing velocity; but Galileo's problem, like that of the quantitative psychologist of to-day, was to make this and other common observations more precise by expressing them in numerical form. There were two simple hypotheses that called for consideration: (i) that the velocity is proportional to distance traversed; (ii) that it is proportional to the time taken. He deduced testable corollaries from each, and by ingenious experiments refuted the first hypothesis and verified the second. From this he went on to disprove a still older hypothesis, namely, that force is necessary to maintain motion (as Aristotle's followers had assumed), and so to show that force is necessary only to change motion. His resolution of the motion of a projectile into two component motions at right angles—one vertical and the other horizontal—in order to test his hypothesis that the path was a parabola—is the earliest instance of factor analysis.¹ Kepler certainly did not neglect the refutation of alternative hypotheses: we are told that he tried nineteen different curves to fit Brahe's observations on the motions of Mars, before he accepted the ellipse. Newton's aim was to find a 'transcendent hypothesis' to co-ordinate these 'empirical generalizations.' Yet he himself formally disproves the chief hypothesis which in those days might have been regarded as a serious rival to his own, namely, Descartes' theory of vortices; and a considerable part of Book II of the *Principia* is devoted to showing that this theory is inconsistent with Kepler's laws.

Moreover, Newton's own statements about the nature of inductive proof appear to recognize the same two points that were emphasized in my own discussion, namely, the desirability of confirming the hypothesis to be established and of confuting the alternative hypotheses. For example, in his letter to Oldenburg (Secretary to the Royal Society) about "the proper method for enquiring into the properties of things," he writes that, in his opinion, the appropriate way to establish a theory is "not by deducing it only from a confutation of contrary suppositions, but also by deriving it from experiments concluding positively and directly" (Macclesfield, *Collection*, II, p. 326).

Since Newton is so often cited in discussions on scientific methodology, it is perhaps worth while glancing at the procedure which he himself actually adopted in one of the more important of his own experimental researches (*Phil. Trans.*, 1672, LXXX, p. 3075). The current theory of light (as expounded, for instance, in Kepler's *Dioptrice*) supposed that pure light was essentially white, and that colour was a "qualification of light" derived from the process of refraction. To test this view Newton began by making a small circular hole in the window-shutter of a darkened chamber, permitting the entrance of a narrow beam of sunlight. This produced a bright circular patch on the opposite wall. He then placed a prism near the entrance. "According to the received laws of refraction," he says, one would have expected the patch still to be "circular," though shifted by refraction from its former position and possibly tinged with colour. Actually the prism produced "a coloured spectrum" of "an oblong form." He then considers whether certain supplementary hypotheses would suffice to explain these peculiarities, e.g., "whether the rays, after trajection

¹ No doubt on many of these points casual anticipations could be found in the work of da Vinci, Stevinus, and others.

through the prism, did not move in curved lines." By varying the size of the hole and the thickness of the glass, he was able to eliminate each one. "The gradual removal of these suspicions," he says, "led me to the *experimentum crucis*." This consisted in placing a second board, containing another small aperture, beyond the prism; and then "turning the prism slowly about its axis." He also subjected the separated rays to a second refraction, showing that each preserved its colour, and that together they could be recombined to give white light. The various results conclusively disproved the corollaries from the accepted theory, and confirmed the corollaries deducible from his own hypothesis, namely, that "light itself is a heterogeneous mixture of differently refrangible rays" and that "the same degree of refrangibility ever belongs to the same colour." It is noteworthy that Hooke criticized Newton's logic on the ground that the explanation reached was still "not the only hypothesis. . . . The same phenomenon will be solved by my hypothesis as well as his." In particular, as is now fully realized, Newton in his discussion about the phenomena of light (e.g., those of polarization) failed to consider that the undulatory hypothesis might take two forms: the waves need not be longitudinal (as Huyghens and Newton assumed); they might also be transverse (as Hooke tentatively suggested).

If Newton failed, it is scarcely surprising that Dr. Richardson should ask: "how can anybody be certain that he has thought of all the possible alternatives?" And Burt, in the chapter that I quoted, also observes that "the difficulty is to be sure that our enumeration of the possible competitors is exhaustive." However, he suggests certain plausible means by which this difficulty may be largely overcome. To begin with, he reminds us that in an inductive enquiry we can never hope to "be certain." We can only strive to reach the most probable conclusion. For this purpose we must begin by "considering the antecedent probabilities of the competing hypotheses." These may differ so widely as to justify us in abandoning any attempt to examine "all the possible alternatives," and in confining ourselves to those that have a high degree of initial probability. Moreover, if a hypothesis (like Dr. Richardson's theological alternative) can be neither confirmed nor confuted, then it is not a hypothesis that the empirical scientist is called upon to consider. Best of all we may be able to "plan the research so that only two alternatives are left open, namely, the hypothesis to be proved and what is commonly called the null hypothesis."

Burt illustrates¹ his argument by deriving (from the familiar rule for compound probabilities) the following formula:

$$P_a = \frac{1}{1 + \frac{p_n q_n}{p_a q_a}}$$

where a denotes the hypothesis to be established, n its contradictory (e.g., the null hypothesis), p the initial probabilities, q the likelihoods after the experiments have been made, and P_a the final or posterior probability. We may assume that neither p_n nor p_a is zero, and (in many cases at any rate) that they do not differ greatly from each other: otherwise we should hardly have embarked on the research. But, except perhaps in Mendelian genetics, they can seldom be determined precisely in any psychological enquiry. If the contradictory hypothesis had the form of a universal proposition (e.g., there are no group factors underlying psychotic conditions), it could conceivably be disproved outright by a single observation; q_n would be zero, and $P_a = 1$. But this is exceptional. Again, if our experiment can be planned so as to verify a corollary that can be deduced logically from hypothesis a , then q_a will be 1; and P_a will depend mainly on the size of q_n . This is the situation that we aim at when we try to establish hypothesis a by disproving Fisher's 'null hypothesis' (i.e., that the results obtained are attributable to chance). It is obvious that the final probability may in some measure be improved either by seeking merely to confirm a (making q_a large) or by seeking merely to disprove n (making q_n small): but the improvement will then as a rule be small, and in general we are not likely to raise P_a much above half unless we seek to do both.

In the particular problem with which Dr. Eysenck and I were concerned, the hypothesis to be proved was that "all forms of psychoses are merely different degrees of one and the same underlying condition, i.e., the differences are quantitative, not qualitative." Here I fancy most psychiatrists would hold that p_n is much higher than p_a . In any case I do not wish to deny that the hypothesis demands that we ought first to seek (as Dr. Eysenck has done) evidence supporting a dominant general factor (thus raising q_a). But that by itself is not enough. In my view it should also be possible to plan a factorial research so that, if there were after all qualitatively different types of psychosis, then we should find corresponding group factors that are statistically significant; while if there were not, then the residuals, even with large samples, would be non-significant: (and in this latter case we should have diminished q_n). I would also add that, to carry complete conviction, the research

¹ In a written note he explains that he regards his formula as illustrative only: "Kneale has argued that the 'probability' of inductions is not the same as the 'probability of chances,' and others have declared that, even if they were, we could not compute them exactly: nevertheless, I hold that the formula may illustrate the way in which what Kneale calls the 'degrees of acceptability' may be lowered or raised: (cf. *Probability and Induction*, pp. 211f., 250f.)."

should test the alternative hypotheses, not only by factorial procedures, but other methods as well ; for, even if a statistical enquiry fails to find qualitative differences, the psychiatrist will still wonder whether clinical, biochemical, or therapeutic methods might not have been successful. In short, as Kneale puts it, "the scientist's aim is to make a theory which fits *all* the facts, and at the same time *has no rival*" (*op. cit.*, p. 253 ; his italics).

I am not wholly satisfied with Dr. Richardson's ingenious way of getting rid of rival theories, namely, by a kind of mutual extermination *a priori*. It would certainly provide a quick solution to the controversy between Dr. Eysenck and myself. Following Dr. Richardson I could argue that "the theory prevalent in the U.S.A." is that all correlations between mental traits are "caused entirely" by group factors ; and that another theory, prevalent in monistic circles, is that they are "caused entirely" by a single general factor. Then I could ask : "do they not refute each other at least so far as the adverb 'entirely' is concerned ?" If so, what is left would be a general-plus-group factor theory, such as I have supported. In this manner the problem would be disposed of straight away, without the need for any empirical evidence whatever ! But would Dr. Richardson be convinced ?

In his own interesting investigation (if I understand his reasoning correctly) he has taken the three hypotheses which seem most plausible *a priori* : then, on deducing their consequences, he has shown that the results of the third seem to agree with facts, whereas those of the first and the second issued in absurdities.

W. L. G.

Editorial Note.—The correspondence on this subject in the previous issue has prompted several letters, asking for further references on the issues raised. On the fundamental question of scientific method, there can be little to add to the full discussions in Russell's *Human Knowledge* (esp. Pt. VI, 'Postulates of Scientific Inference') and the works he there quotes—by Keynes, Nicod, and Broad. Here it is scarcely possible to reply in detail to the questions put by the various correspondents or by the reviewer himself in his opening paragraph. Since, however, he has referred to my own opinions, it may save some misunderstanding if I briefly indicate one or two minor points on which I should disagree.

I cannot help feeling that, in restricting himself chiefly to physicists for illustrations of the logical procedure most appropriate to psychology, he has laid himself open to much the same kind of criticism that he urged against Dr. Eysenck. Jeffreys, in the passage cited by Dr. Richardson, has shown that the methodology of the earlier physicists was by no means so faultless as their psychological imitators have supposed ; and certainly it is hardly the methodology most suitable for the psychologist. For early examples of the kind of reasoning appropriate to psychological experiments, I should prefer to go to biological writers, like Harvey or Cæsalpinus, rather than to physicists like Newton or Galileo.

In the reply to Dr. Richardson it is alleged that "the first instance of the hypothetico-disjunctive method" is to be found, not in Newton (with which I should agree), but in Galileo : much the same statement is to be found in popular introductions and in books by continental psychologists.¹ But as a matter of history, this seems incorrect. So far as European writers are concerned, the revolution in methodology was not so much the dramatic achievement of any single genius ; it was the outcome of a new tradition that had been slowly gathering force, in the face of vigorous criticism, at the University of Padua—a tradition which is traceable back to Buridan of Paris, and so to Buridan's English master, William of Ockam, whose 'razor' implies that hypotheses are more often to be rejected than accepted.² The change itself is in part an outcome of the prolonged disputes between Franciscans and Dominicans, which brought methodological issues to the fore.

As one of the earliest applications of the 'method of exclusion,' I should be inclined to cite, not Galileo's *Demonstrazione Matematiche* (1638), but Harvey's *De Motu Cordis et Sanguinis* (1628).³ The opening sentence of his introduction declares that it will be imperative to begin by stating the

¹ Bertrand Russell, for example, writes : "Scientific method, as we understand it, comes into the world full-fledged with Galileo" (*The Scientific Outlook*, p. 22 : cf. also Lewin, K., *A Dynamic Theory of Personality*, ch. I, 'The Conflict between Aristotelian and Galileian Modes of Thought in Contemporary Psychology'). Just as English writers, from Hume onward, have urged that empirical psychology should take Newton for its model, so continental writers have put forward Galileo. Lewin is of course fully justified in his censures on Aristotle's treatment of *physical* problems ; but that has far less relevance for psychology than psychologists are apt to suppose. In verdict : "Linnaeus and Cuvier may have been my gods, but they were mere schoolboys compared to old Aristotle" (*Life and Letters*, III, p. 252).

² Cf. Duhem, P., *Études sur Leonard de Vinci* (1906-13) and Randall, J. H., 'The Development of Scientific Method in the School of Padua,' *Hist. of Ideas*, I, 1940.

³ 'Le plus beau livre de la physiologie,' as Flourens called it. Harvey and Galileo were at Padua together ; and that University had already produced several revolutionary writers in the biological field, such as Vesalius, Cæsalpinus, Colombo, and Fabricius.

chief alternative suppositions "in order that the false [e.g., Galen's hypothesis of ebb and flow] may be disproved, and the true [viz., Harvey's alternative hypothesis of circulation] may be confirmed, by accurate observation and multiplied experience"; and his unexpected introduction of quantitative proofs into what is a problem in biology rather than physics is a particularly instructive feature of his reasoning.

As regards the explicit theory of inductive methodology, there is ample evidence (in spite of the reviewer's objections) that the change in standpoint was mainly due to Bacon, who in later years was Harvey's patient. Bacon is the earliest writer who definitely insists that "the first duty of induction is rejection or exclusion" (*Novum Organon*, 1620, II, Aph. 16; cf. Sigwart, *Logic*, II, p. 296). Jevons, like most nineteenth century writers, greatly underrated Bacon's early influence. As Reid observes, "the best models of inductive reasoning, such as the 3rd book of the *Principia* and the *Opticks* of Newton, were drawn from Bacon's rules"¹: Newton's very phrase '*experimentum crucis*' is borrowed from the *Novum Organon* (II, Aph. 36). Hume, in the introduction to his *Treatise*, describes the application of the 'experimental method of reasoning,' first to simpler sciences, and then 'at the distance of a century' to 'the science of man,' as having originated in the new principles laid down by 'my Lord Bacon'; and he himself constantly uses the disjunctive type of reasoning.² The psychologist will recall that Bacon announced that his inductive method was to be applied 'non tantum ad naturales, sed ad omnes scientias,' expressly including psychology and social or political theory (I, Aph. 127). Finally, nearly all the eminent scientists associated with the early work of the Royal Society—Boyle, Hooke, Newton, Wren, and Wilkins, together with Petty and Grant, the founders of 'Politically Arithmetic'—acknowledge the lead of the *Artium Instaurator*³: and a contemporary ballad describes how

The Prime Virtuosi have undertaken
Through all the experiments to run
Of that learned Man, Sir Francis Bacon,
Shewing what can, what can't, be done.

For another hundred years a heated conflict dragged on between the English inductive school, which claimed Bacon as its founder, and the French deductive school, deriving from Descartes. But by the middle of the 18th century the inductive method had won; and Bacon is restored to his pedestal by D'Alembert and the *Encyclopédistes*.

Yet once again the formulation of the methods proper to the inductive sciences cannot be credited to any single thinker. It can only be deduced from an historical and critical review of such methods as they have actually been worked out and practised by scientists themselves. Nor is there as yet any final or satisfactory formulation to which the novice can be referred. As Russell has remarked, "the inductive method is the glory of science, but the scandal of philosophy."

In conclusion, may I add that I readily agree with Dr. Richardson that there is still something that the psychologist may learn from Newton's methodology in dealing with the problems of motion, provided we reinterpret his procedure in the light of the later developments to which Dr. Richardson himself alludes? It has, I think, become clear that the so-called 'laws' are not so much 'hypotheses' to be established by inductive or deductive procedures, but simply 'principles' stating how the scientist proposes to deal with the phenomena in question: they express, not the conclusions of his inferences, but the prerequisites of his inferences. When, in dealing with optical phenomena one man chooses as his fundamental concept that of the light-ray, while another chooses that of the light-wave, neither is propounding a generalization which we can show to be either true, or false, or merely probable: each is announcing a working policy he has determined to adopt.⁴ The same is true when Riemann and later Einstein, discarding 'action at a distance,' decide to substitute a 'field theory.'

A recognition of this methodological distinction might, I fancy, go far to ease some of the current disputes in psychology. In the statistical field there are obvious analogues in the use of factor analysis. The fundamental principles preferred by advocates of this or that factorial method are not to be regarded as generalizations inferred from observational data, any more than the principle of the syllogism is a generalization inferred from empirical phenomena; they indicate the kind of

¹ *Works* (1785): p. 200 in Hamilton edition. If a contemporary testimony is desired, I may cite Professor Broad's considered estimate: "In his analysis of inductive arguments Bacon was, so far as I know, breaking new ground; and all later discussion has followed on his lines" (*Selected Essays: Ethics and the History of Philosophy*, 1952, pp. 141-2).

² E.g., *Treatise on Human Nature*, I, iv, 5 ('On the Immateriality of the Soul'): "The absurdity of the last two suppositions sufficiently proves the veracity of the first; nor is there any fourth opinion."

³ Bacon, so named, figures in the frontispiece designed by Evelyn for Sprat's *History of the Royal Society* (1667). The ballad (British Museum) is attributed to J. Glanvill.

⁴ This view has been urged (in somewhat paradoxical terms) by F. P. Ramsey and M. Schlick: cf. also Popper, K. R., *Logik der Forschung*.

simplification each proposes to adopt in constructing a workable map or model of the phenomena he is trying to describe or understand. The choice is rather like that of the maker of geographical maps or globes: according to his purpose he may seek to locate positions on the earth in terms of a three-dimensional sphere, of a curved space of two dimensions, or of an orthographic, stereographic, or Mercator projection. And it would be plainly misleading to treat such devices as general propositions that must be formally proved or disproved by some process of inductive reasoning.

Having decided on the general line of approach and defined the working concepts it entails, we then can proceed to set up and test alternative 'hypotheses' (in the narrower sense of the term); and here I think I am in almost complete agreement with what has been said above about the merits of the 'hypothetico-disjunctive' procedure. For indications as to how precisely it should be employed in any particular research, the reader may be safely referred to Fisher's imaginary 'psycho-physical experiment' as described at the beginning of *The Design of Experiments* (1937, pp. 13-29), or, for more detailed working instructions, to the section dealing with 'Statistical Hypotheses and their Verification' in the recent volume on *Statistical Inference*, by H. M. Walker and J. Lev (1953, pp. 5f.).

C. B.

THE RELATIVE MERITS OF RANKS AND PAIRED COMPARISONS

To the Editor, *British Journal of Statistical Psychology*

SIR,—Recent articles in this *Journal* (Durbin, IV, pp. 85-90; Whitfield, VI, pp. 35-40) suggest that the ranking-methods which were introduced into psychology 50 years ago by Spearman and which were so severely criticized by contemporary psychologists are regaining their old popularity. They are unquestionably attractive to the practical worker; but he would like to be assured that the former criticisms have been adequately met. Titchener and his followers (e.g., Myers and Valentine in this country) maintained that paired comparison was far superior as an experimental technique, since it enabled the investigator to discover preferences, etc., that were psychologically compatible, though logically inconsistent: Brown, in *The Essentials of Mental Measurement*, wrote still more emphatically: "the use of ranks as a measure of a characteristic is inadmissible . . . Some form of frequency distribution must be assumed; the method of ranks assumes it to be a rectangle. The assumption could hardly be more improbable."

Dr. Whitfield and others, who apparently favour the method, appear to disregard these various objections. Moreover, Dr. Whitfield's interesting article in the last issue of this *Journal* raises several additional questions. He shows that the measurements derived from summing m rankings tend, when m is large, to produce a normal distribution. Does not this imply that, if (as Dr. Brown observes) "some form of frequency distribution must be assumed," it would be safer to assume that the distributions are normal? And how are we to calculate the correlation between such a sum and an independent criterion which itself may be given in the form of ranks? Should we re-rank the sums, thus substituting ordinal numbers for cardinal, or should we rather convert the criterion ranks to normal deviates by using Hull's table, as described in so many text-books of psychological statistics?

Dr. Whitfield's discussion of his central problem still leaves many steps somewhat obscure. He mentions two alternative procedures available for testing the statistical significance of a measurement derived from the summation of m ranks, first, that given by Professor Kendall, and secondly, one given by himself and Professor Burt. He somewhat curtly declares that "Kendall's treatment of the problem of m ranks is inappropriate." But he offers no reasons for rejecting this solution or for preferring the second. He quotes Burt as having derived a formula for the frequencies to be expected by chance, but again no proof is given. Since the new solution is of special interest to practical psychologists, could either Professor Burt or Dr. Whitfield insert the missing steps, and could they also explain why the one solution is not appropriate and how the other is deduced?—Yours, etc., R. E. MORRIS.

Reply.—1. The main objections to the technique of ranking were raised not so much by psychologists as by statisticians, notably by Karl Pearson¹; and it is the statisticians, more particularly Professor Kendall, who have restored it to respectability. Brown was unquestionably right in stating that ranks do not form a direct 'quantitative measure' of the quality to which they refer. Ranking is rather to be regarded as a special type of 'manifold classification'—one in which the classes can be arranged in definite order. The use of ties will permit more than one individual to be placed in the same class and so receive the same rank; and, if the appropriate number of ties is forced on the ranker, then his ranks *can* (if we wish) be made to yield any type of distribution, including the normal. But, as Galton pointed out, for statistical purposes an individual's rank-number

¹ Cf. Pearson, K., *Further Methods of Determining Correlation* (Drapers' Company Research Memoirs: Biometric Series, IV, 1907). Brown's criticisms, as he himself observes, are largely a summary of Pearson's. Incidentally, the method of ranks was not "introduced into psychology by Professor Spearman," but by Galton, and popularized by Galton's younger collaborator, J. M. Cattell: (see Guilford, *loc. cit. inf.*, pp. 244f., and refs.).

(or rather his rank-number *minus* one) can best be treated, not as an ordinal number, but as a cardinal number: it states the number of individuals in the classes prior to his.¹ Although from a practical standpoint the distinction may seem somewhat academic, it is nevertheless important for the rigorous deduction of certain formulæ. Thus, with this interpretation of rank-numbers, no assumption need be made about the form of distribution (as Brown erroneously supposed). If so, it follows that, in addition to ease and speed of working, the use of ranks has a special theoretical advantage of its own: it provides a convenient *non-parametric procedure*, i.e., a procedure which can be used when we desire to avoid any specific assumption either about the distribution of the variables or about the scale of measurement. In particular, the analysis of variance takes a very simple form when the data are expressed in terms of ranks.²

2. Let me now deal with the most fundamental of the difficulties raised by Mr. Morris. In his last paragraph he enquires why the 'solution' derived by Dr. Whitfield and myself was preferred to that proposed by Professor Kendall. The explanation is that the two solutions are really solutions to two different problems, namely, the treatment of summed ranks (a) for individuals and (b) for groups.

The difference can be explained more precisely if we take a concrete case. Let us suppose that an education authority has adopted the 'quota system' of allocating children to a grammar school: as a result, School A is allowed (let us say) two pupils, and the teachers are required to name the two brightest individuals out of a batch of eight who happen to have reached the appropriate age. The headmaster accordingly suggests that the four teachers who know all the pupils should each rank them in order of merit, and that the two boys with the highest average rank should be selected.

TABLE I. TEACHERS' RANKS FOR EIGHT BOYS

Boy	Examination	Teachers				Total for Teachers	Average
		i	ii	iii	iv		
T.S.	1	2	2	1	3	8	2
B.R.	2	1	1	3	2	7	$1\frac{3}{4}$
O.F.	3	4	6	8	4	22	$5\frac{1}{2}$
R.L.	4	5	4	7	5	21	$5\frac{1}{4}$
E.B.	5	8	5	6	1	20	5
M.W.	6	7	3	5	8	23	$5\frac{3}{4}$
G.A.	7	6	7	2	7	22	$5\frac{1}{2}$
J.H.	8	3	8	4	6	21	$5\frac{1}{4}$
Sum of Squares of Deviation	(a) about teachers' mean	42	42	42	42	300	$18\frac{3}{4}$
	(b) about boys' mean	$18\frac{3}{4}$	$19\frac{3}{4}$	$27\frac{1}{4}$	$27\frac{1}{4}$	—	—
		93					

Table I shows the orders actually given by the teachers in a case of this sort: the total and average rank for each boy are shown on the right. If we could trust the rankings, then B.R. and T.S. would seem to be the brightest of the eight. It is the statistical investigator's task to determine whether the results *can* be trusted.

Now the theoretical psychologist is interested in the trustworthiness of such rankings obtained in relation to the group as a whole; the practical teacher is interested only in the trustworthiness of the assessments for the two top boys. The questions are plainly distinct.

(a) Let us consider first what is the validity of the rankings *for the whole group*. To determine this, the natural approach will be to carry out a factorial analysis: in that way we can ascertain whether there is any general component common to all four rankings, and, if so, whether it is statistically significant. The correlations between the rankings in Table I, together with the saturations deduced from them, are shown in Table II.

¹ The term 'grade' (a word introduced by Galton) was originally used to designate position as defined by the cardinal number, and distinguished from 'rank,' which is reserved for the ordinal number. Thus Brown, following Pearson, states that "the *grade* of an individual is measured by the number of individuals above him."

² For a fuller discussion the reader may refer to Friedman, M., 'The Use of Ranks to avoid the Assumption of Normality,' *J. Amer. Statist. Assoc.*, XXXII, 1937, pp. 675-701, Kruskal, W. H., and Wallis, W. A., 'The Use of Ranks in One-criterion Variance Analysis,' *ibid.*, XLVII, 1952, pp. 583-621. Cf. also Friedman, M., 'A Comparison of Alternative Tests of Significance for the Problem of *m* Rankings,' *Ann. Math. Stats.*, XI, 1940, pp. 86-92.

TABLE II. CORRELATIONS BETWEEN TEACHERS' RANKINGS

Teachers	i	ii	iii	iv	Total
i	1.0000	.3333	.4048	.2619	2.0000
ii	.3333	1.0000	.2381	.3810	1.9524
iii	.4048	.2381	1.0000	— .0477	1.5952
iv	.2619	.3810	— .0477	1.0000	1.5952
Total ..	2.0000	1.9524	1.5952	1.5952	7.1428 = 2.6726 ²
Saturations ..	.7483	.7305	.5969	.5969	2.6726

To test the significance of the general factor we may follow the procedure outlined in *Factors of the Mind* (ch. X). The average of the observed correlations (i.e., the average of the coefficients in Table I when the self-correlations are excluded) is given (*op. cit.*, p. 275) by the equation

$$\bar{r} = r_{\text{int}} = \frac{1}{m(m-1)} \sum r_{ii'} \quad (i \neq i') \dots (i)$$

$$= \frac{1}{12} \times 3.1430 = .2619.$$

In order to decide whether or not this figure is statistically significant, it is convenient first to calculate a quantity which I denoted by the symbol η^2 . This is given (pp. 275-6) by the equations

$$r = \frac{m\eta^2 - 1}{m - 1}, \text{ or}$$

$$\eta^2 = \frac{(m-1)\bar{r} + 1}{m}$$

$$= \frac{1}{m^2} \sum r_{ii'} \quad (i, i' = 1, 2, \dots, m) \dots (ii)$$

$$= \frac{1}{16} \times 7.1428 = .4464,$$

where $\sum r_{ii'}$ includes the m self-correlations (all taken = 1) as well as the observed inter-correlations. It follows that $\eta^2 = \frac{(\sum r_{ii'})^2}{m^2}$ (p. 276). This quantity is the square of the correlation ratio, and could, if we preferred, be calculated by the more usual formula direct from the original data. We have in fact (p. 274)

$$\eta^2 = \frac{Q_m}{Q_t} \dots (iii)$$

(where Q_m denotes the sum of the squares of the deviations about the means, each squared deviation being taken m times, and Q_t denotes the sum of deviations of each individual mark or rank about the mean for each teacher)

$$= \frac{4 \times 18\frac{3}{4}}{4 \times 42} = \frac{75}{168} = .4464, \text{ as before.}$$

Alternatively, if we write Q_r for the sums of the squares of the residual deviations (93), so that Q_m and Q_r are the 'between groups' and 'within groups' square-sums, then

$$Q_t = Q_m + Q_r = 75 + 93 = 168;$$

and we thus have an alternative method for determining the denominator.

To test the significance of the factor (p. 274) we can now use the procedure adopted for testing the significance of a correlation ratio. Accordingly, we shall take

$$F = \frac{V_m}{V_r} = \frac{Q_m \div (n)_f}{Q_r \div (nm)_f} = \frac{Q_m \div (n-1)}{Q_r \div (n-1)(m-1)}$$

(where V_m and V_r denote the variance 'between groups' and 'within groups' respectively, and $(n)_f$ and $(nm)_f$ denote the appropriate degrees of freedom)

$$= (m-1) \frac{Q_m}{Q_r} = (m-1) \frac{n^2}{1-n^2} \dots (iv)$$

$$= 3 \times \frac{75}{93} = 3 \times .8064 = 2.4192.$$

On referring to Snedecor's table for the distribution of F , we find that this figure is somewhat below the value given for F at the 5 per cent. level. We are therefore forced to conclude that the factor is not statistically significant.

With a larger table we could use a chi-squared test. $m(n-1)\eta^2$ may be regarded as the sum of the squares of a number of standardized deviations; and, as m increases, the distribution of this value tends to that of chi squared. Let us therefore calculate

$$\chi^2 = m(n-1)\eta^2 = \frac{n-1}{m} \sum r_{ii'} \quad (i, i' = 1, 2, \dots, m) \dots (v)$$

$$= 1.250.$$

Entering the chi-squared table with $n-1$ degrees of freedom, we find once again that this value is definitely below that required for the 5 per cent. level.

There are two convenient working formulae that often help to shorten the calculations. To avoid calculating the deviations of the totals (which often introduce fractions), it is quicker to square the sums as they stand, and then correct for the mean. If s_i denotes the total of the rank allotted to the i th boy, then:

$$\eta^2 = \frac{1}{n-1} \left\{ \frac{12 \sum s_i^2}{m^2 n(n+1)} - 3(n+1) \right\} \dots (vi)$$

$$\text{and} \quad r = \frac{1}{(m-1)(n-1)} \left\{ \frac{12 \sum s_i^2}{mn(n+1)} - (3m+1)(n+1) - 2 \right\} \dots (vii)$$

The general procedure outlined above is very similar to that suggested by Kendall¹ in dealing with "the case of m rankings." He proposes to calculate what he terms a 'coefficient of concordance,' expressed by the equation

$$W = \frac{\text{Sum of squares of the deviations of the total ranks about their means}}{\text{The maximum value possible for such a sum}}$$

In my book (*loc. cit.*, p. 275) I pointed out that Q_i , defined as above, is also the maximum value of Q_m . It follows that Kendall's coefficient

$$W = \frac{Q_m}{Q_i} = \eta^2. \dots (viii)$$

He also gives eqn. ii and iv (with W written for η^2); notes that r is equivalent to the intra-class-correlation; and proposes to use Fisher's z -test to test the significance of the variance-ratios. This, of course, is equivalent to the F -test. It would seem therefore that the two methods of dealing with this problem, reached from two somewhat different standpoints, are really identical.

(b) The issue which confronts those who have to decide whether or not this pupil or that deserves a scholarship is quite distinct. They are concerned, not with the statistical significance attaching to the figures for the group as a whole, but with the statistical significance of the figures for some particular individual. In the present example we are required, if possible, to pick out two boys, and two boys only, for a possible scholarship award. According to the summed ranks, B.R. is the best and T.S. nearly as good; but, if we are to justify this choice, we must be able to show that the difference between the figure for T.S. and the average for the whole group is statistically significant. Whether or not the detailed ratings of the other six boys are statistically trustworthy no longer matters.

Consider, then, the total mark obtained by the boy T.S. (who is second in order of merit), namely, 8. This is a deviation of -10 from the average mark, namely, 18. Could such a deviation have arisen by chance?

To decide this question, let us suppose that the teachers have allotted their ranks to T.S. entirely at random. This can best be done by adopting some mechanical method of randomization. We may accordingly imagine that each teacher uses an octagonal teetotum, with its sides labelled from 1 to 8; and, having spun it once, takes whatever number comes uppermost to represent T.S.'s rank. The four random ranks are then summed. Is it likely that a value so high as 8 would thus be secured by sheer chance? In order to obtain an answer, we require first of all to determine the frequency-distributions of such sums when obtained by purely haphazard assignment.

¹ Kendall, M. G., *Advanced Theory of Statistics*, I (1943), p. 411; cf. *Rank Correlation Methods* (1948), ch. VI, and refs. The procedure in my book was intended to be of general application. Kendall has dealt specifically with data in the form of ranks; and, in addition to the more systematic mathematical investigation of this problem, has shown that the procedure may also be used in the case where the ranks are tied.

Mr. Morris enquires how I derived the formula for such frequencies. The problem is an ancient one in the theory of games of chance (see Todhunter, *History of Probability*, 1865, art. 987): it is simply a generalization of the more familiar problem of tossing a coin whose two sides yield two alternative values only. As is well known, when there are only two alternatives, the distribution is given by the *binomial* theorem; when there are more than two, the same reasoning shows that the distribution will be given by the *multinomial* theorem.

Let the number of alternatives (i.e., the number of sides to the teetotum) be n ; let the number of teetotums spun (i.e., the number of teachers judging) be m ; and let s be the sum of the m 'ranks' obtained as a result of the m spins.

Since any one of the n faces may turn uppermost on each of the m teetotums, the number of different ways in which they may fall after spinning will be n^m . To determine the number of ways in which the sum may be obtained, the simplest procedure is to use the device of a "generating function."¹ We find that the number required will be equal to the coefficient of t^s in the expansion of

$$\begin{aligned} & (t^1 + t^2 + t^3 + \dots + t^n)^m \\ &= t^m(1 + t^1 + t^2 + \dots + t^{n-1})^m \\ &= t^m(1 - t^n)^m (1 - t)^{-m}. \end{aligned} \quad \dots (ix)$$

The coefficient of t^s in this expansion will be the coefficient of t^{s-m} in the expansion of $(1 - t^n)^m (1 - t)^{-m}$.

$$\text{Now} \quad (1 - t^n)^m = 1 - mt^n + \frac{m(m-1)}{1.2} t^{2n} - \frac{m(m-1)(m-2)}{1.2.3} t^{3n} + \dots;$$

$$\text{and} \quad (1 - t)^{-m} = 1 + mt + \frac{m(m+1)}{1.2} t^2 + \frac{m(m+1)(m+2)}{1.2.3} t^3 + \dots$$

Multiply the two series together and pick out the coefficient of t^{s-m} in the product. This is easily found to be

$$\begin{aligned} & \frac{m(m+1) \dots (s-1)}{(s-m)!} - m \frac{m(m+1) \dots (s-n-1)}{(s-n-m)!} \\ & + \frac{m(m-1)}{1.2} \cdot \frac{m(m+1) \dots (s-2n-1)}{(s-2n-m)!} - \dots, \end{aligned} \quad \dots (x)$$

the series being continued so long as no negative factors appear in the products. The probability of obtaining s is then obtained by dividing the series by n^m . This, I hope, answers the enquiry made, not only by Mr. Morris but also by several other correspondents who have asked how the formula in the footnote on p. 36 of Vol. VI was actually reached.

In practice, since we shall commonly be concerned with the tail-end of the distribution, there will seldom be more than two or three terms in the series. Here, for example, $s - m = 8 - 4 = 4$; and this is smaller than n . Hence we need consider only the first term of the series; and its value can be obtained from the successive coefficients in the above expansion of $(1 - t)^{-m}$. Their values in turn can be found by taking the 1st, 2nd, . . . figures in the usual table of binomial coefficients (e.g., in *Barlow's Tables*, p. 207), starting with the 1st coefficient for $n = m - 1$. Here, therefore, they are 1, 4, 10, 20, 35. The sum is 70; and dividing this by $n^m = 4096$, we obtain $P = 0.01709$ (as given in Whitfield's table).

Were the numbers large enough for us to assume that the distribution of the sum values was approximately normal, we could argue as follows. The mean of the summed ranks is $\frac{1}{2}m(n+1) = 18$ (as may be verified from the last column but one in Table I); the variance is $\sigma^2 + \frac{1}{12}m(n^2 - 1) = 21$; the absolute value of the deviation (d) will be

$$|\text{sum of ranks for boy} - \frac{1}{2}m(n+1)| - \frac{1}{2} = |8 - 18| - \frac{1}{2} = 9\frac{1}{2}$$

(where the subtraction of $\frac{1}{2}$ is the correction required in the case of a discontinuous variable). Hence $d/\sigma = 2.07$, which gives $P = 0.019$. Thus the attempt to discriminate the two best boys is statistically significant, although the attempt to discriminate the individuals in the group as a whole was not statistically significant. With this particular group the explanation of the divergence is simple: as the headmaster put it in discussing the results, "except for the two lads T.S. and B.R., there is little to choose between the various members of this class, since the duller pupils of that age were left in the class below."

3. Mr. Morris enquires about the procedure to be followed when we have "an independent criterion which is itself given in the form of ranks." As it happens, all the eight boys in the group described above eventually sat for the final examination for junior county scholarships; and a mark-list giving all the candidates in order of merit was circulated to the board of examiners (of whom, as L.C.C. psychologist, I formed one). Accordingly in Table I, I have prefixed a column showing

¹ Cf. Aitken, A. C., *Statistical Mathematics* (1939), pp. 16f.

the ranking of these boys in this later examination; and we may regard this as an "independent criterion" of the kind he has in mind.

Let us therefore suppose that we wish (as Mr. Morris suggests) to calculate some kind of multiple correlation between (a) the composite result of the school teachers' judgements, calculated as an unweighted sum of their several ranks, and (b) the results of the final examination, expressed as an order of merit. How are we to proceed?

Obviously, to convert the sums to ranks, or the ranks to normal deviates,¹ would lower the correlation. The simplest method is to take the unweighted sums as they stand. Let s_i (as before) denote the sum of the teachers' ranks, and c_i the criterion rank; then the ordinary product-moment formula will yield the following equation:

$$r_m = \frac{4 \sum a_i s_i - mn(n+1)^2}{\sqrt{\left\{ \frac{1}{3} n(n^2-1) [4 \sum s_i^2 - m^2 n(n+1)^2] \right\}}} \dots (xi)$$

E.g., with the data given in Table I, we have

$$r_m = \frac{4 \times 732 - 4 \times 8 \times 81}{\sqrt{\left\{ \frac{1}{3} \times 8 \times 63 [4 \times 2892 - 16 \times 8 \times 81] \right\}}} \\ = .748.$$

This formula provides a quick and useful way of reaching an approximate estimate for the multiple correlation with a criterion: (it is actually its lower limit). To find the best value we should have to weight the rankings furnished by each teacher; and these weights would be computed from the correlations and the inter-correlations in the usual way.

To ascertain the improvement that has been effected by basing the correlation on all the m rankings simultaneously, instead of on a single ranking only, we can calculate the average of the m correlations that would be obtained by taking each teacher's ranking singly. There is no need to calculate the correlations explicitly. The average required can be obtained direct by means of the following formula:

$$r_{av} = \frac{3}{n-1} \left\{ \frac{4 \sum a_i s_i}{mn(n+1)} - (n+1) \right\} \dots (xii)$$

E.g., with the figures given above we have

$$r_{av} = \frac{3}{7} \left\{ \frac{4 \times 732}{4 \times 8 \times 9} - 9 \right\} \\ = .500.$$

In this case by taking four teachers' judgements instead of only one we have improved the 'coefficient of prediction,' as it has sometimes been called, by nearly 50 per cent. We can check the accuracy of the foregoing formula by calculating the four correlations separately, and then finding their average in the usual way. We obtain

$$\frac{1}{4} (.524 + .762 + .071 + .643) = .500.$$

Since the distribution of the average rank correlation approaches a normal distribution more quickly than that of a single rank correlation, we may, as a rough test, take the standard error to be

$$= \sqrt{\frac{1}{m(n-1)}} = \frac{1}{\sqrt{28}} = 0.189.$$

The value found is therefore statistically significant.

On re-writing the equation for the multiple correlation (xi) in a form analogous to that of the equation for the average correlation (xii) it will be seen that the numerator is the same in both cases. The denominator differs by the substitution of

$$\sqrt{\left\{ \frac{1}{m} [4 \sum s_i^2 - m^2 n(n+1)^2] \right\}} \text{ for the factor } \sqrt{\left\{ \frac{1}{3} mn(n^2-1) \right\}}.$$

On comparing this with eqn. vi and viii we note that the increase of the multiple correlation over the average correlation depends on the ratio of the standard deviation of the sums of the ranks to the standard deviation of a single set of ranks.

4. Mr. Morris also enquires about the alleged superiority of the method of paired comparison. Carried out in the manner standardized by Titchener,² the procedure is needlessly laborious. Most

¹ For converting ranks into normal deviates I suggest using Table XX of Fisher and Yates's *Statistical Tables* rather than Hull's table, as suggested by Mr. Morris. It may be noted that the correlation between a set of ranks and the corresponding normal deviates is quite high: with a large sample it is $\sqrt{3/\pi} = 0.98$: (a proof is given in the *Notes* previously cited).

² *Experimental Psychology*. I. *Qualitative*. ii. *Student's Manual* (1901), Experiment xxi, pp. 92f.

of its advantages can be secured, and nearly all its disadvantages avoided, if the names of the items to be ranked (in the foregoing example, the names of the pupils) are typed on cards, and each card mentally compared with those nearest it for the quality to be judged. In the case of the eight boys mentioned above, had all the possible comparisons been explicitly made, pair by pair, and the results recorded by each teacher, we should have had a table like that shown below (Table III); my arrangement follows Guilford's (*Psychometric Methods*, 1936, p. 226) rather than Kendall's (*op. cit.*, p. 127). The figures show how many times the boy whose name stands at the top of the column was preferred to the boy whose name stands at the left of the row. In the diagonal I have placed (in brackets) the number of judges making the comparison, not half that number (as Guilford does), because no differences could occur between the judges about the same boy.

TABLE III. TEACHERS' JUDGEMENTS EXPRESSED BY PAIRED COMPARISON

Number of times boy named at top of a column was ranked above boy named at left.

Boys	T.S.	B.R.	O.F.	R.L.	E.B.	M.W.	G.A.	J.H.	Total
T.S.	(4)	3	0	0	1	0	0	0	8
B.R.	1	(4)	0	0	1	0	1	0	7
O.F.	4	4	(4)	2	3	2	1	2	22
R.L.	4	4	2	(4)	2	2	1	2	21
E.B.	3	3	1	2	(4)	3	2	2	20
M.W.	4	4	2	2	1	(4)	3	3	23
G.A.	4	3	3	3	2	1	(4)	2	22
J.H.	4	4	2	2	2	1	2	(4)	21
Total	28	29	14	15	16	13	14	15	144

For data of this kind a suitable method of statistical analysis has been suggested by Kendall (*op. cit.*, ch. XI).

(a) With the full procedure it is usually desirable to begin by testing the judgements for consistency. But, in investigations like those which Mr. Morris has in mind, inconsistencies have no interest in themselves, and merely reveal a defect in the experimental technique: either the quality has not been defined in sufficiently unambiguous terms or the judge is unintentionally departing from the definition; and in that case it would be better to correct the inconsistent judge during the process of judging (allowing ties in case of doubt) rather than to look for inconsistencies afterwards.

(b) To measure the amount of agreement, Kendall proposes a coefficient which he denotes by u . Slightly simplifying his working formula, we may write:

$$u = 1 - \frac{2(m \sum p - \sum p^2)}{\frac{1}{2} m(m-1) \times \frac{1}{2} n(n-1)}, \quad \dots \text{(xiii)}$$

where p denotes the frequencies entered in the triangle below the diagonal only. Thus, with the example given above:

$$u = 1 - \frac{2(4 \times 30 - 54)}{6 \times 28} = .214.$$

But (as Kendall points out) u is in effect a generalization of his own coefficient for measuring the correlation between two sets of ranks, which he denotes by τ . τ , however, is always lower than the ordinary product-moment coefficient of rank correlation (ρ). Hence, to those who are more familiar with the scale of ordinary correlation coefficients, the value computed from eqn. (xiii) may be somewhat uninformative, if not actually misleading.

Since in the cases with which Mr. Morris and I are concerned the inconsistencies will be either non-existent or negligible, I venture to suggest that, for most practical purposes, it will be simplest to calculate η^2 and $\bar{r}$. In experiments where either rank-methods or paired comparison may be used, the relation between the two modes of treatment can readily be seen. On comparing Tables I and III the reader will notice that, although the entries in the body of either table are different, the totals be so. Consider any particular boy, say J.H., the last in the list. According to Table I, Teacher i places him 3rd, and therefore ranks two boys above J.H.; similarly Teacher ii ranks $(8-1) = 7$ boys above J.H.; Teacher iii, three boys; and Teacher iv five boys. The total of such preferences is thus $(3-1) + (8-1) + (4-1) + (6-1) = 21 - 4 = 17$. But this is the number of preferences actually recorded for J.H. in Table III (since the figure 4 in the diagonal does not represent actual preferences). Evidently then, by adding 4, we shall obtain the total entered for J.H. in Table I. Similarly for all the other boys.

It is clear, therefore, that we can use the totals obtained by paired comparison in the same way as we used the totals obtained from the ranks, to calculate the correlation ratio; and from that we may deduce, as before, the average correlation. This method of treating the data has the advantage of keeping the constants so calculated in close relation to the results obtained from ordinary correlational techniques and in particular from a factorial analysis. Hence, except when the method of paired comparison reveals inconsistencies, there would seem to be no advantage in using the coefficient u .—CYRIL BURT.

THE IDENTIFICATION AND INVARIANCE OF FACTORS

By T. LEYDEN

In several papers in this *Journal* (e.g., II, pp. 134f.) various methods have been recommended for deciding how far a factor found in one research may be justifiably identified with that found in another research. The earliest suggestions appear to be those put forward by Burt, who attempted to show that the educational abilities discovered in his later surveys were the same as those reported in the earlier (cf., e.g., *J. Educ. Psychol.*, IX, pp. 68f., and refs.). He suggested (a) the correlation between factor-measurements when the samples are the same, (b) the 'unadjusted correlation' between factor-saturations when the samples are different, (c) a 'symmetry criterion' as an over-all test. More recently Fiske (*Amer. Psychol.*, III, pp. 360f.) and others have proposed (d) the ordinary ('adjusted') correlation between saturations, which has proved extremely popular. Cattell has described (e) a 'symmetrical marker criterion,' and (f) a 'coefficient of pattern similarity' derived from the formula for chi-squared (cf. this *Journal*, II, pp. 136-9, and refs.). In an investigation on which I am engaged it has been surprising to find how divergent, and even contradictory, the verdicts pronounced by these various methods frequently are, especially in the case of the first factor. Many educationists besides myself would, I know, be grateful for a fuller examination of this problem.

To check my own observations, I secured nine sets of test-data, each relating to the same group of tests (or nearly so) applied to the same group of pupils at intervals of one to three years. In three the tests were mixed batteries of intelligence-tests; in six they were educational. Since the tests were not carried out with this particular problem in view, the results deserve only the briefest report. I hope to discuss the details more fully in a later contribution based on a more systematic enquiry planned *ad hoc*.

To compare the results obtained I have taken the correlation between factor-measurements as giving the ideal figure, and have expressed the coefficient obtained from the other procedures as a percentage of this figure.

TABLE I. COMPARISON OF CRITERIA

Criterion	Average Percentage
1. Correlation between factor-measurements	100.0
2. Correlation of factor-measurements with external criterion (teachers' assessments)	87.3
3. Unadjusted correlation between saturations	91.6
4. Adjusted correlation between saturations	14.5
5. Marker criterion	52.7
6. Coefficient of pattern similarity	61.4

The figures above relate only to the first or general factor as extracted by simple summation or the centroid method. With the corresponding 'basic' factor, as extracted by a group factor analysis, the proportions are much the same, except in three cases where there are large supplementary factors: in such cases, not only the relative size and order of the saturations, but also the size of the factor-variances, may appreciably change in the second research. In one case, for example, it appeared that, after an interval of two years, the variance of the general factor obtained by the centroid method had apparently increased from 51 to 54 per cent., whereas the variance of the basic factor had dropped from 46 to 38 per cent. Thus the use of the centroid procedure instead of a group factor analysis may at times obscure the well-known tendency towards increased specialization of abilities with increasing years.

The main conclusion seems clear. With the general factor the popular method of correlating the two sets of correlations may yield a most misleading conclusion: (with bipolar factors, of course, the two coefficients are bound to be much the same). On the whole, the safest method seems clearly to be that based on the 'unadjusted correlation.'

BOOK REVIEWS

Factor Analysis. By Raymond B. Cattell. New York: Harper & Bros. Pp. xiv + 462. £2 8s.

Professor Raymond Cattell's latest book has a threefold purpose, and is divided into three corresponding sections. Part I seeks "to meet the need of the general student in science." Part II is intended to enable the volume to be used as "a text-book for statistics courses." Part III is still more advanced; and, with this further addition, the author hopes that the volume may also serve as a "handbook for the research worker."

As his bibliography shows, Professor Cattell was actively at work on factorial problems nearly a quarter of a century ago—well before other investigators in the United States began to publish contributions in this field. He received his own training at University College, London; and, as a result, unlike most American writers, he not only shows a closer appreciation of the ideas and methods of the earlier British pioneers, but is also able to combine certain features from both modes of approach. His original ideas, his long experience, and his versatile appreciation of a wide variety of devices give his pronouncements on many of the vexed questions of factorial theory an exceptional interest; and an examination of his arguments may therefore in some measure help us to decide on the relative advantages of British and American procedures and theories.

In Part I he sets out the 'Basic Concepts.' Factor analysis is described as a 'wholistic method,' which seeks to discover 'unitary tendencies,' i.e., what are popularly regarded as the 'causes' underlying 'structured groups or clusters' of correlated phenomena. In discussing the "earliest roots of modern factor-analytic ideas," he notes that "both Galton and Pearson had raised the fundamental question"—how can we ascertain and measure the common causes to which covariance is due?; and that both did much to indicate the lines along which an answer might be reached. Further, "a clear statement of multifactor analysis was given by Maxwell Garnett in 1919." Here, as Dr. Cattell incidentally observes, it is "an instructive comment on our understanding of past work" that Garnett's work should "only have been generally recognized after Thurstone had made his discovery." It might be added that Garnett's contribution is not "generally recognized" even now; and even Dr. Cattell seems to have forgotten that Dr. Garnett was Karl Pearson's assistant and explicitly attributes the essential principles to him. Of the various techniques available, we are told, "probably the principal components method—also called the principal axes method—has the greatest appeal to mathematicians." It was in point of fact the first procedure suggested for the purpose, and was developed in a lengthy classical paper published by Karl Pearson as early as 1901.¹ But "the calculations are long and difficult," and the resulting factors appear to possess "no fixed psychological meaning."

In spite of the somewhat laborious computations which any efficient method must involve, the use of some such technique (we are told) is unavoidable, because in the social and psychological sciences the investigator is nearly always forced to deal with several variables—and often with several hypothetical causes—in one and the same research. Professor Cattell therefore lays special stress at the outset on the need for the psychological student to understand the nature of 'scientific method' and the 'proper role of hypotheses.' He is, however, somewhat sceptical of the approbation bestowed by recent factorists, both in this country and abroad, on the so-called "hypothetico-deductive procedure." The logical methods by which science seeks to establish empirical relations, he says, are "essentially those set out by inductive writers like Bacon and Mill." Indeed, in his own view, factor analysis is "far more free than most methods from the necessity to elaborate rigid hypotheses: it is the ideal method of *open exploration*" (his italics), though "of course" (as he says elsewhere) "the factorist always enters with some hypothesis even when he seems to enter with none."

It is partly the desire for freer exploration and partly the belief that the more rigorous methods of Pearson do not yield the 'most meaningful factors' that have led to the supplementary techniques of factor-rotation. Accordingly, 'the art of rotation,' as Professor Cattell describes it, receives full treatment at an early stage. Since we should not expect any one psychological factor to affect more

¹ In Part II, curiously enough, the 'inventor' of the method is said to be Hotelling, who makes no such claim and whose version has been used even less frequently than Pearson's. During the first ten years of the century every writer who investigated the possibility of applying factorial methods to psychological problems—Spearman, Burt, and William Brown—plainly had Pearson's technique in mind, although each would also have endorsed Cattell's criticisms. The fact that Pearson's procedure has not proved practicable or appropriate in psychology without considerable modification should not allow us to overlook that it was after all Pearson who first saw the problem and produced the first multifactor technique.

than a fraction of the total personality, he himself is convinced that the most meaningful scheme will be a 'simple structure' of group factors: such a structure, he considers, will also show most invariance.

How far a factor is invariant must depend, as he rightly points out, on many more conditions than are commonly assumed. Hitherto invariance has usually been interpreted to mean stability of factors, when tests are varied. But the psychologist's initial measurements, like the physicist's, imply variation in three distinct 'dimensions.' As Stern pointed out long ago, they consist of assessments for certain attributes obtained from certain individuals at certain times. Hence it would seem possible, *a priori*, to correlate and factorize either attributes, or times, or individuals. And in some of the earliest British investigations we find profiles derived to describe individual children, and factorizations of test-results obtained on successive days of the week or at different stages of the child's growth. Spearman and Thurstone, on the other hand, seldom mention either the individuals as such or the changes they exhibit with the passage of time; for the most part they and their followers treat factor analysis as concerned primarily with attributes, and chiefly with cognitive attributes. Professor Cattell agrees with those who adopt a wider standpoint. He ingeniously elaborates the box form of Stern's three-dimensional chart, and uses it to provide a somewhat elaborate symbolism to distinguish the various conceivable approaches. Thus "RIB-technique is the typical procedure; Q2B-technique is Burt's advocated use (of correlating persons); Q1S is Stephenson's; and PIB the analysis of the single individual." As Cattell remarks, what might seem at first a complex and confusing situation was greatly simplified by Burt's demonstration of the 'reciprocity principle.' Since in theory the matrix factorized by one method is simply the 'transpose' of the matrix factorized by the other, the factors obtained by one technique must in theory be the same (except for those eliminated by preliminary averaging) as the factors obtained by the other: e.g., to take Cattell's illustration, "if correlation of abilities yields a factor of mechanical aptitude, then correlation of persons will yield a factor or type consisting of mechanically apt persons," and so on. The theorem is really a special instance of Poncelet's 'duality principle,' which has proved so fruitful in projective geometry. This conclusion, although at first vigorously attacked, is, so Cattell considers, fully confirmed by the recent work of Madow. It would seem to follow, as he says, that the views still held by Stephenson about the independence of Q- and R-technique cannot be accepted. In practice of course there will be many minor differences arising from the different modes of approach; but Cattell's final conclusion is that "R-technique provides the binding frame of reference."

Part II deals with 'Specific Aims and Working Methods.' Professor Cattell begins by classifying the various working procedures available. "Of the possible classifications according to the algebraic process, the division into the methods of *simple summation* and *weighted summation* is one which is fundamental for mathematicians and also important for the computer." But there is a further distinction, no less essential to an understanding of the conclusions reached, between what he prefers to call 'bipolar solutions' (i.e., summational and 'centroid' factors) and 'bifactor' (i.e., general and group factor) solutions.¹ The former, he agrees, might be regarded as "difference factors with respect to the latter"; but he doubts whether this is always true. On the whole he considers that "the most likely configuration is one of overlapping group factors with any combination of signs that may be required," in short, a simple multifactor structure which may have negative as well as positive loadings. "The real objection," he says, "to Burt's argument for 'bifactor' methods is that the factors obtained are not the same in meaning as those obtained from simple structure centroid analysis": even if the 'primary abilities' reached by Thurstone—the verbal, spatial, numerical, and other group factors—appear to be similar to those previously found by Burt, there is nevertheless "one factor in the bifactor series, namely, the first general factor, which is *never* matched by anything in the multifactor series" (his italics). But surely this argument will cut both ways. A critic might equally well maintain that the oblique factors obtained in a 'simple structure' are "not the same in meaning" as those obtained by an analysis into general and group factors, and that they omit

¹ This use of terms is likely to be a little confusing to the beginner. The terms 'bipolar,' 'bifactor,' and 'centroid' are so closely associated with the methods of the writers who introduced them that it would seem safer to allow each a kind of copyright for his own specific terms. On p. 133 Cattell appears to suggest that the scheme of general-plus-group factors was a later modification of "the process called bifactor analysis by Holzinger," and was "conceived as a direct extension of Spearman's two factor solution." Spearman, however, explicitly states that Burt's method was "chronologically earlier than the bifactor method" (*Human Ability*, p. 28); and surely the phrase 'bifactor method' is needed to describe the particular working method devised by Holzinger and his co-workers. The resulting factor-matrices may appear the same in their general structure, but in their details the factors and their specifications may differ appreciably. For example, one of Cattell's objections to the 'bifactor solution' is that the results are 'not invariant': this may be true of the results reached by Holzinger's procedure for the reasons he gives; but it does *not* appear to be true of the scheme of factors reached by the so-called 'group factor method.' Indeed, owing to the objectivity of the latter procedure, the results are much more likely to be invariant than those of simple structure, owing to the wide choice of "alternative sets" to which Cattell himself alludes.

the factor which, as a rule, accounts for quite as much of the variance as all the group factors put together. It seems to have been a belated recognition of this point that led Thurstone to suggest re-factorizing the correlated factors, and so securing a 'second order' matrix, containing a general factor as well as group factors.

Cattell's own view of simple structure, it will be noticed, departs in one important respect from that of Thurstone; he does not insist that it should contain positive saturations only. "The myth that a positive manifold must be maintained," he says, "we have already dismissed, since almost any factor is as capable of interfering with a performance as of aiding it." He mentions further difficulties attending the search for simple structure. "The discovery that one and the same correlation matrix may yield more than one simple structure solution, and therefore *alternative sets of factors*, presents a disturbing problem." To secure a decisive result "a common device among the unsophisticated is to rotate for psychological meaning, i.e., to move a factor to a position where the loadings agree with some preconception of the experimenter." But, he continues, "almost any preconception can be confirmed in this way, and this approach merely perpetuates erroneous speculations."

To escape from this situation various devices are considered. One obvious solution, as he points out, is "not to rotate at all! The most systematic defence of this procedure has been made by Burt, followed by Eysenck." But is this a fair representation of these writers' views? Burt, for example, as readers of this *Journal* will be aware, has described both arithmetical and graphical procedures for transforming a set of summational (or 'bipolar') factors to a corresponding set of general and group factors; and, in order to avoid the subjective influences, like those of which Professor Cattell complains, the rotations are based on the dichotomous classification indicated by the bipolar factors. Eysenck, in the earliest of the papers cited by Cattell,¹ follows (as he says) an alternative method "used by Burt and his co-workers nearly twenty years ago"; this is to "make graphs of the [observed] coefficients and judge by the resemblances between the contours": tests that yield parallel contours are then assigned to the same group factor. It is true that there is here no explicit rotation: but the purpose of the calculations was to show that, with certain types of correlational table, what are virtually 'rotated factors' can be reached without "the prolonged and admittedly precarious graphical procedure." In his later work rotation has often been employed.

Elsewhere Cattell himself describes what would seem to be an extension of much the same principle—the use of "parallel proportional profiles." This he would regard as "the ideal method"; but, he adds, "it is not at present in a practicable state," and moreover "for simple structure it has several drawbacks." As a conceivable alternative, he discusses 'criterion analysis,' or, as he prefers to call it, 'criterion rotation'; and adduces "logical objections to regarding it as an independent or sufficient method." In particular, he contends, since it seeks to make the so-called factor "correspond exactly" to the criterion, it is not really a method of factor analysis at all. After a long and instructive discussion of various 'specialized techniques for rotation,' he finally concludes that "bringing the rotation process to a satisfactory conclusion involves more than a mechanical pursuit of simple structure: it requires an appreciation of the meaning of the constellations."

Part III is entitled "General Principles and Problems"; and "is concerned with wider theoretical vistas and with special refinements of practical aids." Here he echoes a plea so often made in the pages of this *Journal*, namely, that, instead of experimentalists and factorists each remaining indifferent or hostile to the complementary approach, efforts should be made to combine the two. What is still more welcome, he puts forward several ingenious suggestions to show how such attempts might proceed.

All Professor Cattell's pages have a racy liveliness that marks the quick and original writer. But some of them at times betray signs of haste. There are slips and obscurities, not only in the phrasing, but also in the figures: e.g., the reference on p. 249 should be '100a,' not '153,' and the reference on p. 315 should be '1a'; the 'correlation matrix' referred to on p. 202 will be found on p. 41, not, as stated, on 'page 46'; on p. 165 question 4 sets the student an exercise in calculating centroid factors, but the 'solution to the question' on p. 166 gives figures that would not be obtained in a centroid analysis, and the factor-variance for the second factor proves to be more than twice as large as that for the first; Table 6 contains a puzzling misprint.

¹ 'Critical Notice of Thurstone's *Primary Mental Abilities*,' *Brit. J. Educ. Psychol.*, IX, 1939, p. 272. A closer study of this paper, which attempts a re-analysis of Thurstone's own correlation table, scarcely bears out Cattell's own conclusions about the superiority of Thurstone's methods. Of the common factor-variance, nearly 60 per cent. can be accounted for by a single general factor, while only 40 per cent. is covered by the seven group factors corresponding to those in Thurstone's "simple structure of primary abilities." Possibly, however, Cattell himself would here agree, since he states that "for some purposes there is little to choose between the two formulations—the rotated centroid yielding primaries and the bifactor yielding a large general factor plus what is left of the primaries."

² The reason is that "the term *criterion analysis* has been already used in a more appropriate sense by other factor analysts."

Inaccuracies and obscurities of this kind, however trivial, are unfortunate in those chapters which provide the student with his first introduction to the working procedure. In Table 1 Professor Cattell sets out the correlation table, which is to serve throughout the book as his 'main example'; and, since his treatment is fundamental to his general exposition, it is worth while examining it in some detail. In ch. III he explains how the student is to extract the first factor by what he calls the 'centroid method,' which, he says, was introduced by Thurstone. Yet both the formulae and the procedure actually followed are, in point of fact, those which had been used by British investigators long before Thurstone introduced the 'centroid method'; and, indeed, Thurstone's modifications do *not* seem to have been adopted. In particular, the method of assessing communalities¹ and the method of reflecting signs depart a good deal from those which are the distinctive features of Thurstone's working procedure.

In the next chapter the student is told how to extract the remaining factors in order of significance until no significant figures are left. But here something seems to have gone wrong either with the account of the procedure or with the way it has been carried out. First, as a little calculation quickly shows, the variance of the second factor is only 0.987, while that of the third is 1.148; secondly, in this last factor the saturations for variables in the negative half of the preceding factor yield positive and negative totals that are far from equal (+.81 and -.65); and finally if we try to reconstruct the observed matrix, the fit is by no means satisfactory. These anomalies appear due to various causes. (1) To begin with Professor Cattell has not noticed that in two places the communalities he has used *destroy the Gramian character* of the correlation matrix, and hence are bound to lead to inaccurate figures. (2) In dealing with what he calls "the exasperating business of reflecting," his own selection of variables for sign-reversal is by no means the best; and it is this that accounts for the low variance of the second factor.² (3) He tells the student that, since the observed coefficients are given to no more than two decimal places, "all the figures in the computation" should be rounded off to two places: but, if only two figures are retained in the computation, the final saturations are likely to show rather large errors in the second place. (4) The method of allocating the signs to the saturations for factors I and II in Table 8 gives *negative* values to those tests which (as even a casual inspection of the original correlation table shows) must have *positive* values in the group factor matrix. It is, therefore, hardly surprising that Professor Cattell is induced to conclude that his centroid factors are devoid of meaning. But of course he might retort that at this stage he is not concerned with meanings, but only with obtaining something to rotate.

In ch. V the student is told 'how to rotate for scientific meaning.' Cattell's Table 9 gives the "orthogonally rotated factor-matrix" reproduced below, and appends new communalities "obtained from the actual loadings by squaring": (incidentally, the squared loadings for test 4 do not add up to the communality stated, namely, 0.98). The reader is then assured that "further rotation fails to improve on the above," and that "the best simple structure has been obtained." Yet on a later page (Table 27) some of the values are improved, without any specific indication of the change. But even so, when the captious reader tries to reconstruct the initial correlations he quickly discovers that there is still room for a good deal of further improvement.

To demonstrate these points, let us examine the results obtained from a straightforward application of the method of simple summation. Let us adopt the method of 'simple summation' in its original form³ and then apply an arithmetical rather than a graphical rotation. One iteration only

¹ I have found more than one of my colleagues baffled by the attempt to discover how the proposed communalities were reached. In the note to Table 1 (p. 41) the student is told that "the method by which they are 'guessed' will be explained later"; but nothing more is said about the matter in this chapter. When he reaches ch. X, he is told that "a rough method—No. 1 below—has already been given in chapter III" ('No. 1 below' consists in taking the highest coefficient in the column, but rounding it off to one significant figure only). But this method has *not* been given in ch. III; and it is plainly *not* the method that has been used in that chapter: (in cols. 4 and 5 "the highest coefficient in the column" is .66, but the figure inserted is .90; in cols. 7 and 8 "the highest coefficient" is .72, but the figure inserted is .80 in col. 7 and .60 in col. 8). A later page suggests rearranging the variables "in order of their mean *r*'s," and considering the general trend instead of just blindly inserting the highest. "These considerations" (it is said) "lead to the practice suggested by Burt," and usually mean "estimating communalities with high mean *r*'s to be higher than their highest *r*'s, and those of low mean *r*'s to be a little lower." This method is called 'No. 2,' and would seem to be the method actually adopted in ch. III; but it is certainly not "that introduced by Thurstone."

² Even without rearranging the figures, an inspection of the table of 'first residuals' seems sufficient to indicate the kind of sign pattern required: by far the largest figures are those in the column for test 5, and here the positive and negative signs plainly contrast tests 2, 4, and 5 with tests 1, 3, 6, 7, and 8. Having decided (presumably from the positive block at the corner) to group 6, 7, and 8 together, it is extremely difficult to see why Dr. Cattell then groups 5 with these.

³ Cf. Burt, C., *Factors of the Mind*, pp. 461f., or earlier papers. The working was carried to four decimal places.

was needed to reach satisfactory communalities; and with these it appeared that all the correlations except the last two could be completely accounted for by three factors only: the last two required a small supplementary factor. The saturations are shown in Table I A. Using a strict arithmetical rotation (based on the procedure described in this *Journal*, III, p. 75) we then obtain the rotated factor-matrix shown in Table I B. These factor-saturations enable us to reproduce the correlations *exactly*. I do not know what may be the source from which Professor Cattell derived his correlation table; but it is tempting to hazard a guess that these were probably the figures used to construct it. The results of Professor Cattell's own rotation are inserted in Table I C. It will be seen that as many as seven of the saturations printed deviate by between .10 and .15 from what may be regarded as the true values: they are deviations which, with a sample of moderate size, might well exceed twice the standard error.

TABLE I. FACTOR-SATURATIONS

Method	A. Simple Summation				B. Group Factors				C. Rotated Centroid Factors (Cattell)		
	I	II	III	IV	I'	II'	III'	IV'	F ₁ '	F ₂ '	F ₃ '
V ₁	.344	-.123	.486	.000	.0	-.1	.6	.0	-.05	-.61	.10
V ₂	.396	.341	.217	.000	.0	.4	.4	.0	.46	-.37	-.04
V ₃	.110	-.384	.401	.000	.0	-.4	.4	.0	-.38	-.47	.13
V ₄	.932	.337	.167	.000	.4	.6	.7	.0	.75	-.57	.38
V ₅	.528	.688	-.384	.000	.3	.9	.0	.0	.85	.08	.17
V ₆	.033	-.133	-.036	.000	.1	-.1	.0	.0	-.11	.01	.14
V ₇	.682	-.385	-.455	.100	.9	.0	.1	.1	.12	.02	.89
V ₈	.613	-.342	-.396	.100	.8	.0	.1	-.1	.10	.04	.80
Sum of Squares	2.276	1.148	.986	.020	1.71	1.51	1.19	.02	1.68	1.06	1.65

In case the reader cares to check the calculations, Table II shows the rotation matrix obtained arithmetically. Since Cattell's bipolar-matrix is based on somewhat unsatisfactory estimates of the common factor-variances for the several tests, it is not possible to obtain a very good approximation to the 'true' values (particularly if an orthogonal transformation is used): I give a transformation based on a least squares fit, and append Cattell's rotation matrix for comparison.

TABLE II. TRANSFORMATION MATRICES

Arithmetical				Graphical			
A. Based on Loadings in Table I				B. Based on Cattell's Loadings			C. Cattell's Rotation Matrix
I'	II'	III'	IV'	I'	II'	III'	F ₃ ''' F ₁ ''' F ₂ '''
I .6873	.3574	.6323	.0000	F ₁ .7093	.3858	.6202	F ₁ .71 .53 .47
II -.4213	.9054	-.0538	.0000	F ₂ .6165	.1236	-.7643	F ₂ .50 .10 .86
III -.5917	-.2293	.7729	.0000	F ₃ .3324	-.9141	.1515	F ₃ .50 .84 .19
IV .0000	.0000	.0000	1.0000				

It will be agreed, I think, that the bipolar-matrix is here by no means devoid of meaning: on the contrary, it clearly indicates the location of the various group factors: the first bipolar (factor II) suggests a group factor with positive saturations for tests 2, 4, and 5 (such as are found in factor II); the second bipolar (factor III) indicates a group factor with positive saturations in tests 1, 2, 3, and 4 (such as are found in factor III); the supplementary factor (factor IV) remains unaltered; and finally factor I suggests a factor with positive saturations in tests 4, 5, 7, and 8, and implies an appreciable overlapping in tests 1, 2, and 3.

Some psychologists will be inclined to ask whether these peculiarities in working methods may not be responsible for some of the rather unexpected factorizations that Dr. Cattell has at times obtained for his own empirical data (cf., for example, this *Journal*, II, pp. 114-30, and Dr. Banks' criticisms and re-analysis in II, pp. 204-18); and possibly a good many will feel that the artificial table thus chosen as the main illustration is scarcely typical of those which the majority of students will encounter in their work. Moreover, if the student possesses some knowledge of mathematics already, he would himself probably wonder why the elaborate calculations, described stage by stage in ch. III, IV, V, X, and XII, are needed for this table. If, in addition, he has already gained experience in group factor procedures, he would omit both the preliminary centroid analysis and the prolonged rotations, and obtain the final figures (those of Table II B) at the very first step.

At the same time it would be unwise to press these criticisms too strongly. Hitherto most text-books on factor analysis have been chiefly concerned with the factorization of human abilities; Dr. Cattell's book is welcome because he has adopted a much broader standpoint and has throughout kept well in mind the factorization of human personality generally. It is therefore conceivable that, in a fuller and more technical discussion, he might be able to adduce instructive arguments in favour of what at first sight seem rather inappropriate procedures.

In any case, whether the reader accepts these methods and whether he agrees with the doctrines expounded or not, it is impossible to deny that the book itself provides one of the most stimulating introductions to the subject, and is full of novel, original, and often highly illuminating suggestions.

E. N. Harper

Tests Collectifs du Centre de Recherches. By M. Reuchlin and E. Valin. Paris: Institut National d'Étude du Travail et d'Orientation Professionnelle. 1953. Pp. 152. 400 fr.

This monograph presents a study of the battery of tests compiled by the Research Centre of the *Institut National d'Orientation Professionnelle*. The first part summarizes the main practical conclusions; the second and longer part deals with the more theoretical questions of methodology. Here the authors have "devoted an important place to certain procedures which seem of special interest for such work, but have nevertheless been but little used in France," namely, the procedures usually called 'factor analysis,' "to which," they say, "our analysis appears to make some contribution."

"It was," they explain, "between 1910 and 1920 that Burt found that correlations between cognitive tests could most frequently be accounted for by two types of factor—a general factor common to all and a series of group factors each common only to a limited number: this hypothesis was opposed by Spearman, who at first refused to admit the existence of group factors; but it is equally opposed to the analysis into 'simple structure' developed later on by Thurstone." However, M. Reuchlin and M. Valin take the view that, in the latter case at any rate, the apparent conflict is by no means irreconcilable; and they consider that their own figures clearly demonstrate that "the results obtained by Thurstone's method (provided it is pushed to its final conclusion) lead to those obtained more simply by the method of analysis already described by Burt." It would be instructive to hear how far Thurstone and his followers accepted this conclusion.

For the purposes of their main enquiry a battery of 15 tests was administered to nearly 2,000 boys and girls aged approximately 14 years. The children were divided into four subgroups—town boys, country boys, town girls, and country girls. A table of correlations was calculated for each group and then factorized by various procedures. Twenty-six pages of tables give their results in some detail.

The writers begin with an ordinary 'bipolar analysis' based on the principle of 'simple summation.' This leads them first of all to discuss the similarities and differences between Burt's 'simple summation method' and Thurstone's 'centroid method.' The most important divergence, in their view, arises from the way the reduced self-correlations (or 'communalities') are assessed: Burt's method is to estimate the self-correlations by inspecting the general trend of the observed correlations (not relying solely on the largest coefficients), and then to correct these provisional estimates by the results obtained from an actual analysis, repeating the whole process until stable figures are reached. Thurstone, on the other hand, simply inserts the largest figure from each column, both when extracting the first factor and when dealing with the later factors. The French investigators consider that the procedure adopted by British workers "is much more satisfying in principle, but obviously more exacting to apply." With the present data they find about four iterations usually sufficient; and their final calculations, they believe, warrant the extraction of four factors.

The bipolar factor-matrices reached by this procedure they regard as somewhat difficult to interpret, because "the negative saturations cannot easily be justified psychologically: in the cognitive field it is not easy to imagine two groups of tasks of such a nature that success in one group necessarily leads to relative failure in the other." This argument, however, assumes that the factors obtained are to be identified, just as they stand, with cognitive abilities. But according to Burt—obtained as they have earned wide acceptance—factors are primarily principles of classification, and the view has earned wide acceptance—factors are primarily principles of dichotomous classification. On this basis the results can

at once be interpreted. In three of their groups, the first bipolar factor evidently contrasts tests requiring abstract or intellectual processes with those requiring more concrete or practical processes; in the fourth group the same division is effected by the second bipolar factor. In each case the other bipolar factor contrasts contents—namely, verbal with numerical and spatial with logical.

To secure an interpretation in terms of positive abilities, the authors next proceed to 'rotate' the initial factor-saturations, first in accordance with the methods described by Thurstone and secondly in accordance with those originally proposed by Burt.

In Thurstone's procedure, the first stage is to obtain a 'bipolar-matrix' of 'centroid factors.' The second stage consists in applying a succession of graphical rotations until some kind of 'simple structure' is achieved, which will consist of group factors only with no general factor. The investigators offer very convincing evidence for regarding the results so reached as inadequate if not actually misleading. It is true that, by permitting 'oblique' or correlated factors, Thurstone usually succeeds in avoiding anything like a general factor; but this, it is argued, leaves readers quite in the dark about the extent to which the various tests chosen for the battery really tend to measure the same thing: in short "they are kept in the same state of ignorance, and therefore exposed to the same kinds of error, as in the days when factor analysis had not been developed, and test-results were commonly interpreted by mere inspection." Thurstone, we are told, in a desire to simplify the naïve scheme of general and group factors (i.e., of 'general ability' and 'special aptitudes'), attempts to throw the entire onus of explanation on group factors or special aptitudes only, just as Spearman at one time hoped to explain everything by appealing to a general factor only, i.e., the factor of 'general intelligence.' But, it is argued, since the general or basic factor contributes quite as much to the total variance as the group factors do, the new type of 'parsimony' seems just as indefensible as the old. Why, simply on the grounds of personal preconceptions, should we ignore the useful general factor any more than we ignore the narrower group factors? The conclusion therefore which the authors draw is that "whenever oblique factors are used for the first order analysis, a second order analysis is absolutely indispensable."

At this third stage, as they very justly assert, it is not sufficient to calculate just *one* general factor of second order; the saturations for *all* the second order factors in all the tests should be computed. Unfortunately "no American author using second order factors has, to our knowledge, pushed the analysis as far as this, although the necessary technique permitting this has been described by Thurstone." At this stage it will once again be necessary to begin by estimating the diagonal values or 'self-correlations'; for this purpose, they say, "nous avons estimé d'abord les communautés par 'smoothing,' suivant la méthode décrite à propos de l'analyse par 'sommation simple' de Burt, et ensuite procédé à des réestimations successives." The final outcome of the Thurstone analysis consists in four factor matrices—one for each batch of children: they include, besides the first or 'basic' factor, three supplementary in the case of the town boys and town girls, four with the country girls, and five with the country boys. In the last case the additional factor really forms a second general factor; its appearance is attributed to the fact that (as not infrequently happens with Thurstone's procedure) the correlations between the oblique factors would lead to self-correlations over 1.000, if a single general factor alone were used. In the main the supplementary group factors thus obtained are very similar in nature to those which appear in the 'simple structures': but they have the advantage of being orthogonal instead of oblique or correlated. The inter-correlations both of the tests and of the first order factors are now expressed—and better expressed—by a general factor. It is consequently inferred that "analyses into first order oblique factors, which dispense with a general factor, cannot of themselves provide a satisfactory foundation either for theory or for practice."

The writers turn next to what they call the 'méthode de Burt.' In the main they follow the procedure outlined in the article on 'Group Factor Analysis' in this *Journal* (III, pp. 40-75). They agree that this procedure is far less subjective than that of Thurstone, but contend that the subjective element is still not wholly eliminated. British factorists commonly assume that, if a summational analysis is carried out first and the resulting sign-pattern adopted as a guide (to be superseded of course by the results of successive approximation), then the final subdivision of the tests or traits into groups and subgroups emerges automatically. Drs. Reuchlin and Valin feel that, owing to the lack of any test of statistical significance, there may still be room for arbitrary decisions, when the factor-saturations are extremely small. Moreover, in calculating the preliminary matrix of non-overlapping group factors, they are a little dismayed to find that somewhat different cross-classifications appear with the different groups of children; for instance, two of the tests involving number are classed sometimes with the tests requiring numerical ability and sometimes with the tests of reasoning. But, since each test involves both processes, and since boys and girls (or urban and rural pupils) may conceivably make use of either process in varying degrees, it is not surprising to find such discrepancies; indeed, so far from constituting defects of the method, they probably indicate genuine differences in the way different pupils tend to attack such problems.

Where there is any doubt, however, the full factorial procedure should, if possible, be followed, and then every decision will be determined by the data. If, for example, during the initial stages

a low negative saturation leads to an incorrect grouping, this will be mechanically corrected when the non-overlapping factors are replaced by overlapping factors. No doubt in practice these elaborate precautions are seldom employed. Unless the numbers tested are extremely large, the minor changes that would result could, as a rule, claim only moderate probability, and consequently it seems legitimate to shorten the procedure, as indeed the French investigators have done. Where a good many iterations are required, they point out that the procedure can be greatly facilitated, "en procédant, comme Burt le fait souvent, par condensation," or by graphical devices. But probably most psychologists would hold that, whenever the doubtful points are of importance, the wiser course would be, not to spend time on more refined calculations, but to plan a new and more suitable experiment to solve the issues so raised.

At the later stages, where rotations are required to show the extent of overlapping, the French writers prefer (with most British investigators) to rely mainly on an arithmetical method of rotation. For the final figures, however, instead of continuing the iterative procedure when only slight modifications are likely to result, they have "effected a few slight orthogonal rotations with the graphical procedure." In doing so, they are inclined to wonder whether their procedure may not be "une utilisation maladroite," departing a little too far from the rigorous ideal. But, as Burt and others have observed, provided the rotations are carried out 'blindly' and in accordance with perfectly objective principles, graphical procedures may be used (by those who prefer them) even when the computer wishes to obtain a matrix of general and overlapping group factors which are perfectly orthogonal. Though seldom so accurate, they are occasionally speedier than arithmetical iterations.¹

The results at which the investigators finally arrive are highly instructive. All the factor-matrices clearly reveal the tendency towards "the recurrent pattern of g , v , u , and k " which so often appears in British work. In the tables obtained from the country boys, there is first a general or basic factor accounting for nearly 70 per cent. of the common factor-variance (rather over 40 per cent. of the total variance) and four group factors—verbal, numerical, spatial, and (the smallest of all) reasoning. Moreover, in these instances there is practically no overlapping. In two groups, however, the reasoning factor shows a tendency to fuse with the numerical and the spatial factors. In the results obtained from the country boys and from the country girls, there is a large group factor covering the 'intellectual' tests (as distinct from the 'practical'), and this then tends to subdivide into separate verbal and numerical factors which do not overlap.

But the finding on which the writers lay greatest stress is the remarkable similarity in every group between the factor-matrices obtained with Burt's relatively simple methods and those eventually reached with Thurstone's second-order analysis when carried through all four stages to its logical conclusion. The two methods therefore do not really lead to inconsistent results as is so often supposed: a 'reconciliation' is possible. Further, "il nous paraît important et équitable de souligner que cette réconciliation n'a pas exigé que Burt modifie sa conception première: ce sont ses 'contradicteurs' qui se rangent, ou semblent devoir se ranger, à son avis."

P. Durant

¹ In such cases the graphical method adopted is not quite the same as that used by the French authors. Instead of the first arithmetical rotation, based on non-overlapping group factors, the computer usually begins with a preliminary arithmetical rotation by a standard 'transclassificatory operator' (This is a special type of 'equilibrated orthogonal matrix': for its nature and use, see this Journal, I (1947), p. 8.)

It may be noted that, in what would seem to be the earliest published example of factor-rotation, Garnett adopted a simple graphical procedure (*Proc. Roy. Soc., A*, XCVI, pp. 91f.). Burt also used it in factorizing the first large American investigation on the 'Correlation of Mental Abilities' (Simpson, B. R., *Columbia Contributions to Education*, 1912). Indeed, it was Spearman's subsequent criticisms of the alleged group factors found in this table (*Abilities of Man*, 1925, pp. 145-6, 151) that demonstrated the need for a more rigorous procedure or establishing such factors.

INDEX TO VOLUME VI

INDEX OF AUTHORS

<i>Authors</i>	<i>Titles</i>	<i>Page</i>
Burt, C.	Scale Analysis and Factor Analysis	5-24
Burt, C.	The Relative Merits of Ranks and Paired Comparisons	112-119
Dunsdon, M. I., and Roberts, J. A. Fraser	The Relation of the Terman-Merrill Vocabulary Test to Mental Age in a Sample of English Children	61-70
Ehrenberg, A. S. C.	On Making Statistical Assumptions	41-42
Eysenck, H. J.	The Scientific Study of Personality	44-46
Gourlay, N.	Covariance Analysis and its Applications in Psychological Research	25-34
Guttman, L.	A Note on Sir Cyril Burt's Factorial Analysis of Qualitative Data	1-4
Leyden, T.	The Identification and Invariance of Factors	119
Morris, R. E.	The Relative Merits of Ranks and Paired Comparisons	112
Richardson, L. F.	The Submissiveness of Nations	77-90
Richardson, L. F.	On the Obligation to Refute Rival Hypotheses	107
Roberts, J. A. Fraser, and Dunsdon, M. I.	The Relation of the Terman-Merrill Vocabulary Test to Mental Age in a Sample of English Children	61-70
Scarr, Elizabeth H.	Changes in Terman-Merrill I.Q.s with Dull Children	71-76
Sen, Amya	A Preliminary Study of the Thematic Apperception Test	91-100
Slater, P.	The Factor Analysis of Negative Correlations	101-106
Whitfield, J. W.	The Distribution of Total Rank Value for One Particular Object in m Rankings of n Objects	35-40
Wrigley, J.	The Publication of Tables	52

INDEX OF SUBJECTS

<i>Titles</i>	<i>Authors</i>	<i>Page</i>
Covariance Analysis and its Applications in Psychological Research	Gourlay, N.	25-34
Factors, The Identification and Invariance of	Leyden, T.	119
Hypotheses, On the Obligation to Refute Rival	Richardson, L. F.	107
Nations, The Submissiveness of	Richardson, L. F.	77-90
Negative Correlations, The Factor Analysis of	Slater, P.	101-106
Personality, The Scientific Study of	Eysenck, H. J.	44-46
Qualitative Data, 'A Note on Sir Cyril Burt's Factorial Analysis of'	Guttman, L.	1-4
Ranks and Paired Comparisons, The Relative Merits of	Morris, R. E.	112
Ranks and Paired Comparisons, The Relative Merits of	Burt, C.	112-119
Rank Value, The Distribution of Total, for One Particular Object in m Rankings of n Objects	Whitfield, J. W.	35-40
Scale Analysis and Factor Analysis	Burt, C.	5-24
Statistical Assumptions, On Making	Ehrenberg, A. S. C.	41-42
Tables, The Publication of	Wrigley, J.	52
Terman-Merrill I.Q.s, Changes in, with Dull Children	Scarr, Elizabeth H.	71-76
Terman-Merrill Vocabulary Test, Relation of, to Mental Age in a Sample of English Children	Dunsdon, M. I., and Roberts, J. A. Fraser	61-70
Thematic Apperception Test, A Preliminary Study of the	Sen, Amya	91-100

BOOKS REVIEWED

<i>Authors</i>	<i>Titles</i>	<i>Page</i>
Brower, D., and Abt, L. E.	Progress in Clinical Psychology	55-56
Burt, C.	Contributions of Psychology to Social Problems	57
Burt, C.	The Causes and Treatment of Backwardness	58
Cattell, R. B.	Factor Analysis	120-125
Ferguson, T.	The Young Delinquent in his Social Setting	56-57
Fisher, R. A.	Contributions to Mathematical Statistics	53-54
Reuchlin, M., and Valin, E.	Tests Collectifs du Centre de Recherches	125-127

